LIBRARY
UNIVERSITY OF MIAMI

Max F Meyer

LIBRARY

THE UNIVERSITY OF MIAMI · FLORIDA · MCMXXV
Magna Est Veritas

LIBRARY
UNIVERSITY OF MIAMI

STATISTICAL METHOD

TEXT-BOOK SERIES

EDITED BY PAUL MONROE, PH.D.

TEXT-BOOK IN THE HISTORY OF EDUCATION.
By PAUL MONROE, PH.D., Professor of History of Education, Teachers College, Columbia University.

SOURCE BOOK IN THE HISTORY OF EDUCATION.
FOR THE GREEK AND ROMAN PERIOD.
By PAUL MONROE, PH.D.

PRINCIPLES OF SECONDARY EDUCATION.
By PAUL MONROE, PH.D.

TEXT-BOOK IN THE PRINCIPLES OF EDUCATION.
By ERNEST R. HENDERSON, PH.D., Professor of Education and Philosophy, Adelphi College.

DEMOCRACY AND EDUCATION. AN INTRODUCTION TO THE PHILOSOPHY OF EDUCATION.
By JOHN DEWEY, PH.D., Professor of Philosophy, Columbia University.

STATE AND COUNTY SCHOOL ADMINISTRATION.
SOURCE BOOK.
By ELLWOOD P. CUBBERLEY, PH.D., Professor of Education, Stanford University, and EDWARD C. ELLIOTT, PH.D., Professor of Education, University of Wisconsin.

STATE AND COUNTY EDUCATIONAL REORGANIZATION.
By ELLWOOD P. CUBBERLEY, PH.D.

THE PRINCIPLES OF SCIENCE TEACHING.
By GEORGE R. TWISS, B.SC., Professor of the Principles and Practice of Education, Ohio State University.

THE PRUSSIAN ELEMENTARY SCHOOLS.
By THOMAS ALEXANDER, PH.D., Professor of Elementary Education, George Peabody College for Teachers.

HOW TO MEASURE IN EDUCATION.
By WILLIAM A. MCCALL, PH.D., Assistant Professor of Education, Teachers College, Columbia University.

A HISTORY OF THE FAMILY AS A SOCIAL AND EDUCATIONAL INSTITUTION.
By WILLYSTINE GOODSELL, PH.D., Assistant Professor of Education, Teachers College, Columbia University.

THE EDUCATION OF WOMEN.
By WILLYSTINE GOODSELL, PH.D.

STATISTICAL METHOD.
By TRUMAN L. KELLEY, PH.D., Professor of Education, Stanford University.

A HISTORY OF EDUCATION IN THE UNITED STATES.
By PAUL MONROE, PH.D. *In preparation.*

STATISTICAL METHOD

Max F Meyer

BY

TRUMAN L. KELLEY, PH.D.

PROFESSOR OF EDUCATION IN STANFORD UNIVERSITY

New York

THE MACMILLAN COMPANY

1923

All Rights Reserved

PRINTED IN THE UNITED STATES OF AMERICA

HA
29
.K4

COPYRIGHT, 1923,
By THE MACMILLAN COMPANY

Set up and electrotyped. Published, May, 1923.

LIBRARY
UNIVERSITY OF MIAMI

PREFACE

This book has been written with a view to serving two needs; that of biologists, economists, educators and psychologists, who know little of higher mathematics, possibly care less, and who use statistical methods merely as a device to portray the facts of their group investigations; and that of those in the same fields who resort to mathematics to aid in the discovery of new truths.

The elementary statistical needs in the four fields mentioned seem to me to be the same and it is my aim to meet those needs and provide a foundation which will serve for advanced work in any one of them.

The approach to the essential principles developed is through concrete problems, only varying from this where simplicity of problems or the necessity for conserving space warrants.

In order to provide a rigorous foundation for further statistical research — which would immediately take the economist, educator, or psychologist as well as the biologist into the fertile field developed by Karl Pearson and his co-workers — the notation follows that of the English school, making such simplifications as are possible for the immediate problems, but endeavoring at no time to introduce a symbol, an approximation, or a lax proof which would have to be unlearned in undertaking more advanced work. The statistician cannot fail to note that the sheer visual weight of symbol, so appalling to the tyro, has been genuinely reduced by the introduction of a few new symbols in connection with multiple correlation.

The fields represented by various correlation and other measures whose probable errors are unknown has been treated very succinctly. I can see no value except at times a slightly greater ease of manipulation, in using a measure whose probable error cannot be calculated if one with a known probable

25987

error and serving the same purpose exists. I have, therefore, simply included and defined such measures for those desirous of using them, without deriving or attempting to justify them.

I particularly request the critical analysis by fellow statisticians of my determinations of probable errors, and such charity in reporting shortcomings as may be due one who has acted upon the policy that as shrewd an estimate as possible of the probable error of a statistical constant is better than no estimate at all. The derivation of probable error formulas has been one of the most difficult undertakings of this text and I cannot expect that the results are faultless.

My statistical training has been rather desultory and it has occasionally been impossible for me to give due credit to the discoverers of well known formulas.

I would, however, say that my greatest inspiration has been the product of that master analyst, Karl Pearson, and that the English school entire has been most contributive. My greatest indebtedness to men in America is to my teachers, Henry Lewis Rietz and Charles C. Grove, for enlightenment upon theoretical points and to Edward L. Thorndike for suggestions as to problems in need of statistical analysis.

T. L. K.

CONTENTS

CHAPTER PAGE

I. THE TABULATION AND PLOTTING OF SERIES 1

SECTION

1. Introduction 1
2. Statistical Series 2
3. Construction of Statistical Tables 5

II. GRAPHIC METHODS 9
4. The Histogram and Frequency Polygon . . . 9
5. The Time Chart; Relative Time Chart; Chart of Ratios 16
6. Smoothing Data 27
7. The Ogive Curve 31
8. The Growth Curve 34
9. The Graphic Representation of Categorical Measures 37

III. THE MEASUREMENT OF CENTRAL TENDENCIES 44
10. Averages 44
11. The Arithmetic Mean 45
12. The Median 54
13. Percentiles 57
14. The Mode 60
15. The Harmonic Mean 63
16. The Geometric Mean 65
17. Weighting 67

IV. MEASURES OF DISPERSION 70
18. The Mean Deviation 70
19. The Quartile Deviation 75
20. The 10–90 Percentile Range 75
21. The Standard Deviation 77
22. The Standard Error of the Mean 82
23. The Standard Error of Any Moment 84
24. The Standard Error of a Class Frequency; of the Median; and of a Percentile 86

CHAPTER PAGE

V. THE NORMAL PROBABILITY DISTRIBUTION . 94

SECTION

25. Derivation of Equation of Normal Distribution . 94
26. Certain Properties of the Normal Distribution . 95
27. Kelley–Wood Table of the Normal Probability Integral 97
28. Further Properties of the Normal Distribution . 98
29. Properties of Portions of a Normal Distribution . 99
30. The Probability of Exceeding a Given Divergence 102
31. Summary of Facts Concerning the Normal Distribution 104

VI. COMPARABLE MEASURES 109

32. The Conditions Requisite for Comparison . . 109
33. The Ratio Method 110
34. The Standard Measure Method 114
35. The Equivalence of Successive Percentiles Method 118

VII. THE FITTING OF CURVES TO DISTRIBUTIONS 123

36. Methods of Fitting Curves to Observations . . 123
37. The Principle Underlying Pearson's Method of Curve Fitting 124
38. Description of Types of Curves 128
39. The Fitting of the Most Important Types of Curves 136
40. The Bearing of Curve Type upon Stability of Distribution 138
41. Illustrations of Unstable Distributions . . . 146

VIII. MEASURES OF RELATIONSHIP 151

42. The Problem of Concomitant Variation in the Sciences 151
43. Findings Resulting from Galton's Graphic Treatment 153
44. Algebraic Statement of Galton's Graphic Findings and Derivation of Correlation Formulas . . 156
45. The Detailed Steps in the Calculation of Correlation and Regression Constants 161
46. The Error Involved in Certain Approximations . 164
47. The Bearing of Broad Categories upon Correlation 167
48. Properties of Correlation Surfaces 172
49. Standard Deviations and Correlations of Various Constants 175

CHAPTER PAGE

VIII. MEASURES OF RELATIONSHIP—*Continued*

SECTION

50. Formulas for the Calculation of the Product-Moment Coefficient of Correlation 179
51. The Interpretation of Regression Coefficients . . 181
52. Product-Moment Correlation of Non-Rectilinear Data 185
53. The Rank Method of Calculating Correlation . 191

IX. FUNCTIONS INVOLVING CORRELATED MEASURES 196
54. Correlations of Sums or Averages 196
55. The Reliability Coefficient 200
56. Correction for Attenuation 204
57. Reliability of Averages 205
58. The Probable Error of a Coefficient Corrected for Attenuation 208
59. Estimates of True Scores and the Probable Errors of These Estimates 212
60. Accuracy of Placement on Basis of a Single Score 214
61. Average Intercorrelation 217
62. The Effect of Different Ranges upon Correlation of Similar Measures 221
63. The Effect of Different Ranges upon Correlation of Different Measures 223
64. The Effect of Double Selection upon Correlation of Different Measures 228

X. FURTHER METHODS OF MEASURING RELATIONSHIP 231
65. The Various Ways of Measuring Relationship . 231
66. The Median Ratio Correlation Coefficient . . 231
67. Correlation Determined from a Curve of Correspondence by Rank 234
68. Correlation Ratio Method 238
69. Method of Parabolic Regression 245
70. Bi-Serial r Method 245
71. Bi-Serial Eta 249
72. Tetrachoric Correlation 253
73. Correlation in a Four-Fold Point Surface . . . 259
74. Measures of Correlation not Equivalent to the Product-Moment Coefficient; Yule's Coefficients of Association and of Colligation 260

CHAPTER PAGE

X. FURTHER METHODS OF MEASURING RELATIONSHIP — *Continued*

SECTION

75. Measures of Relationship Interpreted in Terms of Probability 262
76. Equi-Probable r 265
77. Mean Square Contingency and Coefficient of Contingency 265
78. Variate Difference Method 271

XI. MULTIPLE CORRELATION 279
79. The Problem 279
80. Theoretical Treatment — 3 Variables 280
81. Three-Variable Problem Illustrating Meanings of Constants 285
82. The Use of the Alignment Chart 291
83. The General Treatment of the n-Variable Problem 295
84. The Method of Successive Approximations . . 302

XII. STATISTICAL TREATMENT OF SUNDRY SPECIAL PROBLEMS 311
85. Statistical Constants Determined from Mutilated Distributions 311
86. Correlation Determined from Mutilated Distributions 314
87. The Probable Error of Percentage Measures of Overlapping 316
88. A Criterion for the Addition or Elimination of Elements Having Fixt Weightings 319
89. Trade Test Calibration 320
90. The Determination of the Cross-over Value of a Chromosome Section 321
91. The Best Weighted Average of Independent Variables 324
92. Psychophysical Methods 326

XIII. INDEX NUMBERS 331
93. The Bearing of Purpose and Material upon Form of Index 331
94. The Meaning of a Price Ratio and of a Price Index 333
95. The Probable Errors of Various Indexes . . . 334
96. The Accuracy and Flexibility of the Weighted Geometric Mean Index 339

CHAPTER PAGE

XIII. INDEX NUMBERS — *Continued*

SECTION

97. Criteria for Judging of the Excellence of Indexes . 341
98. The Use of Any Year as Base 346

APPENDIX

A. LIST OF IMPORTANT SYMBOLS 349
GREEK ALPHABET 356

B. BIBLIOGRAPHY 357

C. KELLEY–WOOD TABLE OF THE NORMAL PROBABILITY INTEGRAL 373

INDEX 387

ALIGNMENT CHART . . . . . *inside back end-paper*

STATISTICAL METHOD

CHAPTER I

THE TABULATION AND PLOTTING OF SERIES

Section 1. Introduction

Two occasions for resort to statistical procedure, the one dominated by a desire to prove a hypothesis, and the other by a desire to invent one, have led to two schools of statisticians.

The first school is that represented by mathematicians who start with certain elementary principles and deduce therefrom facts of distribution, frequency and relationship. In so far as observed situations parallel these conclusions the same elementary principles are supported as applying to the data in hand. One weakness of this approach lies in the fact that a number of causes — different sets of elementary principles — may result in substantially the same net result. A still greater weakness is that it is essentially a deductive procedure and relatively sterile in suggesting new causes — in inspiring creative inferences. It is fundamentally a method of proof and not one of invention; and just because it is a method of proof, it has a permanent place in statistical method. It must, however, if in the service of the social and biologic sciences, be but a handmaid to the creative genius of mathematical analysis and induction.

The second school is best represented by those biometricians and economists who start with observed data and endeavor so to group them and treat them that the constant features of the data are made apparent. This is a process of statistical analysis. It may at times be expected to be an involved process, for social phenomena are complex. Data are frequently warped to fit statistical convenience, but if statistics is to realize its high destiny, procedure must be flexible, for only when the method is mobile can it fit immobile data. The accurate

measurement of those features of phenomena which are exceptional is the unique province of statistical analysis.

The method of approach in this text is inductive, starting with data and deriving constants, and will not give the noumenal satisfaction that comes from tossing coins, throwing dice, and sorting cards, thus obtaining distributions which approach an ideal standard.

Mathematical statistics form very much of a unit, and it is impossible to treat fully of topics in an order which does not call in earlier chapters for concepts developed later. The genuine unity of statistics is made apparent by these interrelationships, and I have not attempted to avoid them. Terms used in an earlier part of the text than that in which derived are usually unambiguous on account of the context, but should there be any difficulty in understanding, the reader is directed to the bold face references given in the index and to the list of mathematical terms and symbols given in the Appendix.

Section 2. Statistical Series

The treatment of this and the succeeding section largely follows that of Day (1919 and 1920).

A statistical series is a succession of facts having some common characteristic. A series may be thought of as either giving (1) a location in time, (2) a location in space, (3) an indication of qualitative difference, or (4) of quantitative differences.

(1) Trends in prices, rates of growth, fatigue, learning and forgetting curves, diurnal changes, etc., are illustrations of the magnitude of a variable with reference to time. Temporal series have certain characteristics which necessitate a technique in their interpretation which is peculiar to them. Any time series of appreciable duration (in studying etheric vibrations .001 of a second would be a very appreciable duration) may be expected to show periodic fluctuations. As a consequence one of two procedures is necessary, dependent upon whether (*a*) it is desired to study the changes within a certain cyclical period, or (*b*) to study trends independent of such periodic changes. Illustrations will make the problem clear:

(*a*) Let it be required to ascertain the nature of the load of an electric power generating plant during a twenty-four-hour period. The current consumed per hour for some one day could be tabulated or plotted. The result would have only such accuracy as would result from a single day's sampling. To obtain a more reliable picture, a number of days could be combined and the tabulation made showing the average load for each hour of the twenty-four. Obviously error might creep in here, for the load on a Monday would be quite different from that on a Saturday or Sunday and perhaps different from that on the other days of the week. With due allowance for holidays, probably a very satisfactory idea of the hourly fluctuations of the Monday load could be obtained by pooling results for several Mondays. Differences in daylight, temperature, etc., would make it unsound to combine all the Mondays in the year. The problem cited is typical of temporal series problems and the principle that should guide one in pooling results should be to group as wide a range of data as are typical with respect to the characteristic under investigation, but not affected by other seasonal or systematic tendencies.

(*b*) Let it be required to ascertain the nature of the seasonal fluctuations of the load. In this case a tabulation by weekly units would be the best as this would completely suppress both Saturday and Sunday and hourly idiosyncrasies. With this in mind it is seen that a tabulation by six or eight day or monthly periods would not be as satisfactory as weekly or bi-weekly periods. The principle to follow is to use such a temporal unit as equals or is an integral multiple of the period within which occur the tendencies which it is desired to suppress.

A second characteristic of a temporal series arises from the general lack of significance of the absolute value of a function at a given time. Interpretation depends upon the relation of the function at one time to its magnitude at a second time. This fact has led to the use of index numbers, or ratios of magnitudes. The magnitude at a stipulated time is considered basic and used as the denominator of all the ratios. The index number is not limited to temporal series, but it is more characteristic and more generally serviceable with them than with

other series. Many considerations enter into the choice of the base, but if there is one time, such as a certain year, which more than any other shows a constant condition of the function, or an ideal or desirable condition, it will have special value as the base.

(2) Just as index and periodic concepts are fruitful in interpreting temporal series, so is the map essential in portraying spatial series. Many spatial series show both qualitative and quantitative differences, in which case considerable ingenuity is needed to devise a map with cross sectioning, or color scheme, to portray the essential facts. Spatial series are intrinsically more amenable to graphic treatment, and less to numerical treatment, than temporal or quantitative series. The maps of the U. S. Coast and Geodetic Survey, of the Weather Bureau, and of the Census Bureau show the completeness, variety and detail of portrayal possible. The groupings of territories in spatial series and the subdivision of areas may follow conventional procedure or the peculiar needs of the problem. The order adopted by the Census Bureau in giving population statistics is as follows:

TABLE I

New England	West North Central (*continued*)
Maine	Missouri
New Hampshire	North Dakota
Vermont	South Dakota
Massachusetts	Nebraska
Rhode Island	Kansas
Connecticut	South Atlantic
Middle Atlantic	Delaware
New York	Maryland
New Jersey	District of Columbia
Pennsylvania	Virginia
East North Central	West Virginia
Ohio	North Carolina
Indiana	South Carolina
Illinois	Georgia
Michigan	Florida
Wisconsin	East South Central
West North Central	Kentucky
Minnesota	Tennessee
Iowa	Alabama

TABLE I (*continued*)

East South Central (*continued*)
- Mississippi

West South Central
- Arkansas
- Louisiana
- Oklahoma
- Texas

Mountain
- Montana
- Idaho
- Wyoming

Mountain (*continued*)
- Colorado
- New Mexico
- Arizona
- Utah
- Nevada

Pacific
- Washington
- Oregon
- California

(3) Qualitative series are those in which the classification is based upon the presence or absence of certain qualities. They lead to categorical distributions and are treated statistically by means of the probabilities of frequencies, and by measures of relationship dependent upon the same — contingency coefficients, etc. The variability of a frequency is the basic concept in the statistics of qualitative series.

(4) Quantitative series are those in which the classification is based upon the degree to which some measured trait is present. They are the most amenable to numerical treatment and their consideration comprises the bulk of this text. The variability of a distribution is the most basic concept in the statistics of quantitative series.

Life's problems do not confine themselves to single series, and certain methods have been developed for handling problems which are complexes of two or more of the four types mentioned, but it is well to recognize that in general the problem and the method are functions of a single series.

Section 3. Construction of Statistical Tables

The chapter which follows this deals with graphic methods and is concerned with charts, diagrams, graphs, etc., constituting pictorial representations of statistical series. The statistical table is quite different. Its purpose is not directly to give a picture of a sequence, but to provide the basic data from which such a picture, or at least the outstanding features of such a picture, may be determined and visualized if desired.

The statistical table is simply a shorthand statement of facts. If a thousand or so facts of the sort, "The population of Aaber County is 4000;" "The population of Anthony County is 3200;" "The population of Avery County is 4800;" etc., etc., are to be presented, they can not only be more concisely shown by tabulation, but several thousand additional facts, such as "The population of Anthony County is 800 larger than that of Aaber County" are presented at the same time and in an agreeably compact manner. The desire to accomplish double, triple, or manifold presentation by a single tabular arrangement is the desideratum which imposes conditions and determines appropriateness of procedure.

The same facts in regard to population are shown in the following five tables, and while not exhausting the possibilities of presentation these will suffice to show the wide option which exists in presenting very simple data.

TABLE II

Populations and Areas of Counties

Counties	Population 1920	Area in Sq. Miles
Aaber	4,000	480
Anthony	3,200	400
Avery	4,800	800
Bascomb	16,000	700
Brown	3,000	600

TABLE III

Areas and Populations of Counties

Counties	Area in Sq. Miles	Population 1920
Aaber	480	4,000
Anthony	400	3,200
Avery	800	4,800
Bascomb	700	16,000
Brown	600	3,000

TABLE IV

Counties arranged according to Population

Counties	Population 1920
Brown	3,000
Anthony	3,200
Aaber	4,000
Avery	4,800
Bascomb	16,000

TABLE V

Counties arranged according to Population

Counties	Population 1920
Bascomb	16,000
Avery	4,800
Aaber	4,000
Anthony	3,200
Brown	3,000

TABLE VI

Counties arranged according to Population

Population 1920	Counties
16,000	Bascomb
4,800	Avery
4,000	Aaber
3,200	Anthony
3,000	Brown

As judged by a single purpose no two of the tables given are equally meritorious. If the table is to be used more frequently in abstracting information about various counties than as a means of comparing counties, i.e., if it is a reference table and not one pointing some conclusion, the items in the stub (the first column) should be arranged alphabetically as in Tables II and III in order to facilitate the finding of items desired. If populations are more likely to be studied than areas, Table II is preferable to Table III, as the Population column holds a dominant position in Table II.

Should it be intended that the table be not primarily a reference table arranged to simplify the extraction of items of information, but, let us say, to point conclusions with reference to populations, Tables IV, V, or VI are preferable to Tables II or III. If counties of large population are the chief consideration, Table V is preferable to Table IV, as the first row of a table ranks higher in dominance than successive rows. Next in importance is the last row. Totals or averages are, because of their importance, frequently placed in the first row, but if other items demand this position or if captions (headings of columns) are less readily interpreted when separated from the body of the table by a row of totals or averages, then the bottom row may be used.

As a means of pointing conclusions dependent upon populations Table VI is to be preferred to Tables IV or V, as the population data hold the dominant position in Table VI.

In general one should so draw up the table that the items in the stub and the captions constitute the argument or information with which the table is entered, and so that the column and row next to the stub and captions contain the most important items to be obtained from the table. Rows and columns more removed from these dominant positions should contain less important data, except that the last row and last column may be given to data of first or second importance.

Such Tables as II and III are primary or general purpose tables, since they contain the raw data without abridgment, and may be used for various purposes. Such Tables as IV, V, and VI are derived from primary tables, such as II and III, and by emphasizing certain facts serve a special purpose.

These two types of tables should be recognized. The special purpose table is always published because it conveys the point of the study. The general purpose table should always be published also, as it provides the only means of checking the author and of discovering if other or further conclusions can be drawn. Several tables and many calculations may be involved between the primary and the final derived table. If full description of these intermediate steps be given it is not essential that these intervening tables and calculations be published.

CHAPTER II

GRAPHIC METHODS

Section 4. The Histogram and Frequency Polygon

The picturing of facts, when the nature of the data permits, conveys a readier comprehension than is possible from any array of figures. The accurate graphic portrayal of data is therefore the problem of this chapter.

Since there are but two dimensions to the surface of a sheet of paper, ordinarily but two series of facts are shown in a single graph. Consider the accompanying data giving the maximum temperatures recorded by the Weather Bureau for each day in July and August, 1917, for New York City.

TABLE VII

Maximum Temperature for Each Day

July 1–Aug. 30, 1917

N. Y. City

July		Aug.			
1	80	1	98	17	85
2	88	2	96	18	80
3	74	3	83	19	81
4	78	4	80	20	84
5	81	5	82	21	85
6	80	6	82	22	80
7	79	7	88	23	76
8	70	8	78	24	83
9	75	9	83	25	82
10	65	10	80	26	74
11	66	11	82	27	82
12	71	12	83	28	80
13	81	13	83	29	83
14	81	14	78	30	81
15	75	15	81	31	75
16	85	16	80		
17	87				
18	80				
19	77				
20	83				
21	81				
22	86				
23	86				
24	86				
25	84				
26	85				
27	90				
28	80				
29	81				
30	95				
31	98				

TABLE VIII

Tally Sheet

Temperatures	No. of Days with Given Temperature	Temperatures	No. of Days with Given Temperatures
65	|	82	|-|-|-|
66	|	83	|-|-|-| | |
67		84	| |
68		85	| | |
69		86	| |
70	|	87	|
71	|	88	| |
72		89	
73		90	|
74	| |	91	
75	| |	92	
76	|	93	
77	|	94	
78	| |	95	
79	|	96	|
80	|-|-|-| |-|-|-|	97	
81	|-|-|-| | |	98	| |

If it is desired to study diurnal changes in maximum temperatures a graph could be made in which the abscissa (the horizontal dimension) represents the days in order, July 1, July 2, etc., and the ordinate (the vertical dimension) represents the temperatures in order, 0°, 1°, 2°, etc. For July 1 the ordinate would be 80, for July 2, 88, etc. A line connecting the successive ordinates would give a picture of the changes in maximum temperature throughout the two months. Or, it may be desired to disregard the sequence of the days and obtain a general idea of what constitutes the maximum temperatures for days in New York during July and August. In this case the abscissa will represent temperatures and the ordinate the number of days. To do this, Table VIII is first made out from the data in Table VII and then plotted as shown in Charts I or II.

Chart I is a histogram or a pictorial representation by means of rectangles, telling precisely the same story as a table of frequencies, such as Table VIII. Chart II is a frequency polygon. It is not a series of discrete elements as are the raw, gross, or original, measures, but a closed figure, each part of

which is connected with the next, giving the idea of continuity in the measures. Each of these graphic forms has its advantages; the histogram in case heights of rectangles are to be accurately compared; and the frequency polygon if the idea of continuity is desirable. Note that in drawing the frequency polygon points *a* and *c* are connected and not points *b* and *c*

CHART I

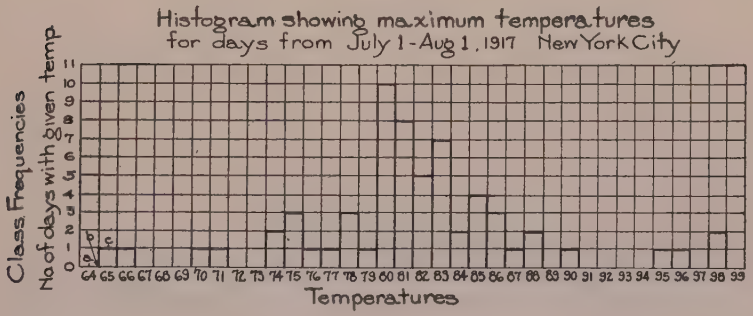
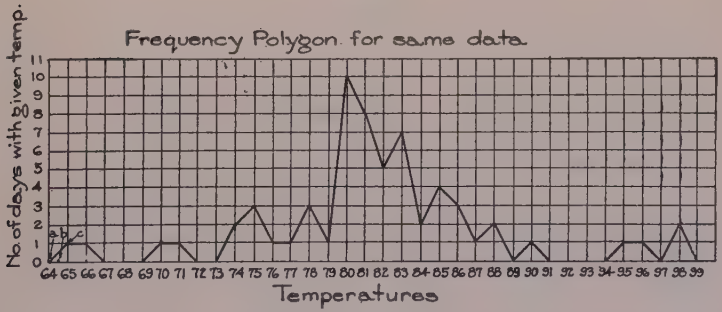

Histogram showing maximum temperatures
for days from July 1 - Aug 1, 1917 New York City
Class Frequencies
No of days with given temp
Temperatures

CHART II



Great care should be taken to insure that the graph agrees with the labels of the coördinates. Note that the class index "65" designates the mid-point of the interval, the lower limit of which is 64.5 and the upper limit 65.5; that in the polygon, point *c* is directly above the class index 65, and that in the histogram the class index 65 designates the mid-point of the horizontal dimension of the rectangle.

It is allowable to label the beginning and end of the interval. In such case the histogram or polygon would be drawn exactly as given and point *b* would be labeled "64.5" and under no circumstances "65."

It has become somewhat customary in educational fields to speak of a child as solving 10 problems in a speed test, meaning thereby that 10 problems were solved and the 11th started but not finished when time was called. In plotting the distribution of scores the designating number, 10, has been placed at the beginning of the interval. No objection should be made to this were the numerical computations in harmony with this procedure, but very generally such scores have been treated as exactly 10.0 in calculating arithmetical averages with the result that the curve and the constants computed from the data do not agree. Not uncommonly such scores have been treated as 10.0 scores in calculating means and as 10.5 scores in calculating medians, with the result that a comparison of mean and median scores gives an entirely erroneous impression as to the skewness of the data. This faulty procedure has probably been followed unwittingly, but unfortunately with the sanction of teachers. The following is quoted from page 50 of the Second Year Book — Division of Educational Research, Los Angeles, July 1919:

"LESSON SIX — THE ARITHMETIC MEAN

Method of Finding the Mean

No. Problems	No. Pupils	
12	3	$3 \times 12 = 36$
11	5	$5 \times 11 = 55$
10	7	$7 \times 10 = 70$
9	4	$4 \times 9 = 36$
8	2	$2 \times 8 = 16$
	21	213

213 divided by 21 equals 10.14 the mean. The median in the same distribution would be 10.64." In this lesson problem the mean is in error if 12 implies the interval 12.0 to 13.0 and the median (see Section 12) is in error if it implies the interval 11.5 to 12.5. The error here cited probably grew out of an error in labeling a distribution. Uniformity is needed, and it would be in harmony with well-nigh universal procedure in

the physical and biological fields to consider a score of 10 as being also a class index, or mid-point of an interval. Should this lower the grade of a few million school children by one half a point no harm would be done and the great advantage of having the recorded test score measures exactly those to be used in calculating means, standard deviations, correlations, etc., and of having the recorded measures also the class indexes in graphs is attained. Throughout this text a score no matter how derived originally is uniformly to be interpreted as covering an interval extending from half a unit below to half a unit above. The accompanying data provide a nice problem in plotting where the distribution is decidedly asymmetrical; where a part of the distribution is lacking; where the class intervals (i.e., range covered by successive groups) are unequal; and where the existence of a few excessively extreme measures makes it impossible to select coördinates (abscissas and ordinates) which satisfactorily reveal the entire distribution.

TABLE IX

British Income-tax Payers — 1914
American Consular Report, May, 1915

Income	No. of Assessments
£ 160 to 200	257,499
200 300	237,434
300 400	85,557
400 500	46,063
500 600	23,411
600 700	13,383
700 800	10,250
800 900	5,779
900 1,000	7,445
1,000 2,000	16,363
2,000 3,000	3,381
3,000 4,000	1,231
4,000 5,000	678
5,000 10,000	882
10,000 and over	390
	709,746

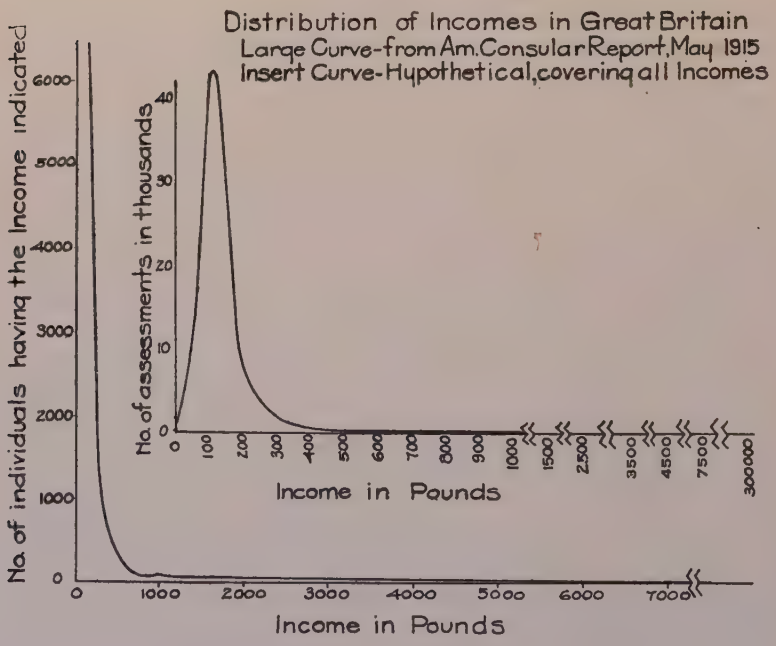
TABLE X

Income	No. of Assessments
£ 0 to 40	150,000
40 80	750,000
80 120	1,680,000
120 160	1,400,000
160 200	400,000
200 300	390,000
300 400	97,000
400 500	49,000
500 600	24,000
600 700	14,000
700 800	10,000
800 900	6,000
900 1,000	7,000
1,000 2,000	17,000
2,000 3,000	3,000
3,000 4,000	1,000
4,000 5,000	700
5,000 10,000	900
10,000 and over	400

Notice that the first class interval covers a range of £40 while the next to the last extends over £5000 and that the last interval extends over an amount not recorded but probably

as large as £100,000. No scale which will satisfactorily picture the £40 class interval will be satisfactory for a £100,000 interval. The curve below (not the insert curve) pictures as much of the distribution as possible. Even with an interval of £1000 to a distance of one-half inch, space does not permit of showing the last interval. Having omitted this class it is necessary to make note of the fact as has been done in the lower right hand corner of the chart.

CHART III



Since the first interval is £40, the second £100, the tenth £1000 and the fourteenth £5000 it is impossible to plot ordinates proportionate to the frequencies: 257,499; 237,434; 16,363; and 882; and truly picture the situation. Some account must be taken of the difference in size of intervals, for the ordinate should represent the number of cases per unit interval. Accordingly 257,499 has been divided by the interval represented, 40, giving 6437, the number of persons per range of £1; 237,434

divided by 100, giving 2374, etc., which quotients are the heights of the ordinates representing the respective classes.

The ordinates have been joined by a smooth line to emphasize, even more than does the frequency polygon, the idea of continuity. A polygon or histogram is generally to be preferred, as it is less likely to be misleading.

Having the data of Table IX for incomes above £160 it is possible to make a sufficiently close estimate of the total distribution of wealth in Great Britain as to suggest what the major features of the actual distribution would be. Let us therefore assume the total distribution of wealth to be as recorded in Table X and investigate its salient features.

The plot of the data of Table X is given in the insert. Since the abscissa scale is much larger than before it has been impossible to plot the entire distribution without breaks. These breaks are indicated, as should always be the case, by prominent pairs of zig-zag lines. Note that the ordinates, which were obtained as before, are plotted at the mid-points of the intervals, e.g., there are 390,000 individuals receiving incomes from £200–£300, or 3900 per £1 of the range. This ordinate, 3900, is erected at £250, the class index and also the mid-point of the interval.

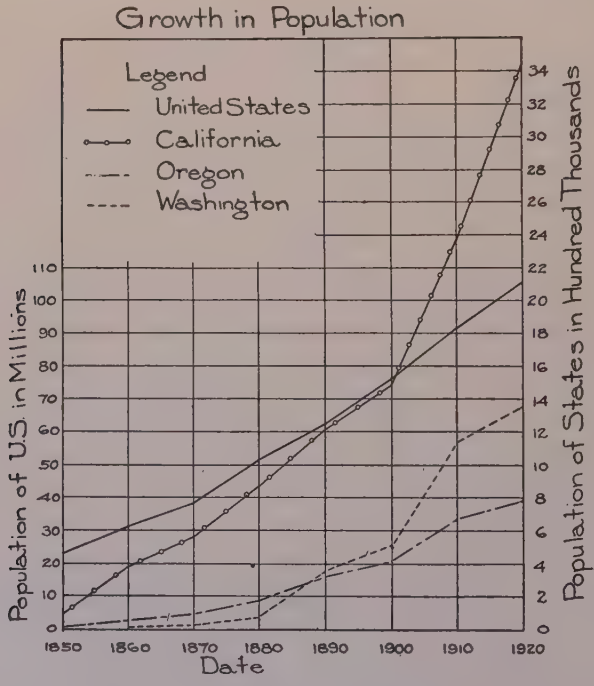
The shape of the curve indicates that there were more than 3900 per £1 for incomes between £200 and £250 and less than this number per £ for amounts between £250 and £300.

It may also be noted that since a curved line connects the points, the area lying under the curve and between £200 and £300 will not total exactly 390,000 as it should. In curves smoothed by visual inspection such inaccuracy is practically unavoidable. For these particular data a frequency polygon would be still less satisfactory as it would indicate a mode at £100 whereas, assuming the hypothetical data to be correct, the mode is somewhat above that amount. A histogram would give the most accurate presentation, but would be less satisfactory in other respects. The total area in a given histogram interval is accurate, but the rectangular distribution within the interval indicated by the histogram may be quite inaccurate if the interval is large.

Section 5. The Time Chart; Relative Time Chart; and Chart of Ratios

Charts have been presented in which the ordinates were frequencies and the abscissas amounts in a gross score. Such graphs are ordinarily characterized by small frequencies at either end of the distribution and a single mode somewhere in between. If, however, frequencies are plotted as ordinates, and periods of time as abscissas, a different type of curve is found, for generally with the passage of time the function continues to grow or at least persist. The following data and chart are characteristic:

Chart IV



Note that the right hand axis is labeled from the bottom up. Simplicity and clearness can frequently be obtained by labeling the lines in a chart and omitting the legend.

TABLE XI

Population in Thousands

	1850	*1860*	*1870*	*1880*	*1890*	*1900*	*1910*	*1920*
CALIF.	93	380	560	865	1,213	1,485	2,378	3,427
ORE.	13	52	91	175	318	414	673	783
WASH.		12	24	75	357	518	1,142	1,357
U. S. ENTIRE	23,192	31,443	38,558	50,156	62,948	75,995	91,972	105,711

The graph shown illustrates the use of a single set of abscissas and two sets of ordinates for the plotting of two kinds of curves upon the same chart; (1) population of the United States in millions and (2) population of States in hundred thousands. This method is usually very misleading and the present illustration is no exception. Double ordinate charts can be used with less error if, going with changes in time there are changes in the general direction of the curve, i.e., if it rises and falls, for then if a second curve also showing such fluctuations in direction of trend is plotted on the same chart it is possible to compare the one with the other as to direction of fluctuation, but it is not possible at all accurately to compare them as to magnitude of fluctuation. The method should be used with very great parsimony and precaution.

For the chart shown the comparisons which can validly be made are those of absolute growth between state and state. The curve for the entire United States confuses rather than helps in the comparison. Absolute growth in the United States cannot be compared with absolute growth in the states as the scale is 1/50 that used for the states. Relative growth in the United States and in the states cannot be determined by comparing the slopes of the curves — e.g., the slope of the curve for the United States between 1900 and 1910 is steeper than that for Oregon for the same years, but the percentage growth for that period for the United States is $21\left(\frac{91972 - 75995}{75995} \times 100\right)$ which is less than the percentage growth for Oregon, $63\left(\frac{673 - 414}{414} \times 100\right)$. Likewise it is ap-

parent that relative growth of state and state is not shown by these graphs.

The Relative Time Chart

Relative growth could be shown by plotting the populations for the several years in terms of some one year as a base, or "relative." For the data in hand this would be unsatisfactory for no matter what year is taken as the relative (e.g., 1850, . . . 1910, 1920) the resulting graph would be difficult of accurate and significant interpretation. If change over a short period only is under consideration, relative curves reveal significant tendencies, especially if the measures, in particular the base measure, are large with respect to fluctuations.

The following data permit of portrayal in graphs, either in terms of original scores or as ratios.

TABLE XII

Chicago Data *

Year	U. S. Entire Dunn's Wholesale Price Index	Av. Yearly Retail Price Round Steak	Union Wage per Hour		
			Painters	Linotype Operators	Carpenters
1907 . .	107.264	14.3¢	50¢	50¢	56 3¢
1908 . .	113.282	14.9	50	50	56.3
1909 . .	111.848	15.9	55	50	56.3
1910 . .	123.434	16.2	60	50	60
1911 . .	115.102	15.9	60	50	60
1912 . .	123.438	19.1	60	50	65
1913 . .	120.832	20.2	65	50	65
1914 . .	124.528	22.3	70	50	65
1915 . .	124.168	21.2	70	50	65
1916 . .	137.666	22.6	70	50	70

* U. S. Dept. of Labor, Bur. of Labor Statistics. Union Scale of Wages and Hours of Labor, 1916.

Chart V is a graph of the data of Table XII and Chart VI of Table XII a. In Chart V there are various breaks in the vertical scales permitting the use of three different sets of values. The location of the word "Date" in Chart VI is preferable to that in Chart V.

TABLE XII a

(Prices and wages expressed as ratios, 1907 as base)*

Chicago Data

Year	Dunn's Wholesale Price Index	Average Yearly Retail Price Round Steak	Union Wage per Hour			Retail Relative Price
			Painters	Linotype Operators	Carpenters	22 Common Articles of Food
1907 . .	100	100	100	100	100	100
1908 . .	106	104	100	100	100	105
1909 . .	104	111	110	100	100	109
1910 . .	115	113	120	100	107	113
1911 . .	107	111	120	100	107	113
1912 . .	115	134	120	100	115	121
1913 . .	113	141	130	100	115	120
1914 . .	116	156	140	100	115	124
1915 . .	116	148	140	100	115	124
1916 . .	128	158	140	100	124	138

* The decimal point is omitted, as usual, so that a ratio of "106" means a six per cent increase.

CHART V

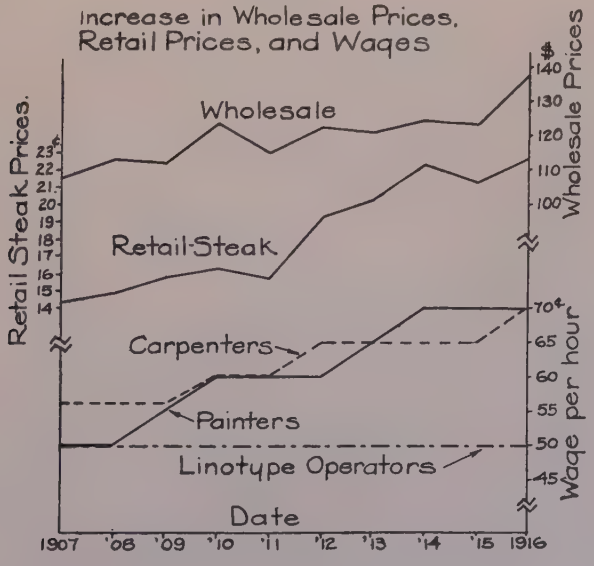
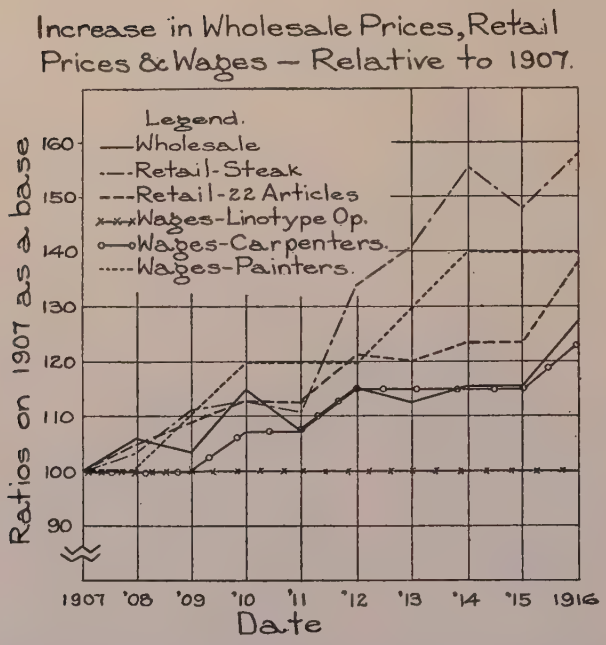


CHART VI



Neither of the accompanying graphic presentations is without serious drawbacks. From Chart V it is possible to infer that the retail price of round steak and wholesale prices of food products both dropped from 1910 to 1911 but it is not possible to judge which suffered the greatest relative decline. Chart VI does show that relative to 1907 wholesale prices suffered most.

Chart VI gives the impression that painters are better off than carpenters, — relative to condition in 1907 they are, but in no other sense as Table XII shows. A relative table or chart shows facts relative to condition at date of base and nothing else, which is a point that must be stressed or it will be overlooked by the untrained reader. A gross measurement table, or chart, reveals gross changes and directions of relative changes but not the magnitude of relative changes.

Another inaccuracy which is commonly present in ratio measures and accordingly in charts based upon them, is due

to the fact that variations in ratios are frequently large with respect to the base used. Prices may increase or cities grow 101, 200 . . . 1000 per cent, but it is impossible for them to decrease by such amounts. A change in ratio from 50 to 100 means more than a change from 100 to 150 though they show up the same when plotted. Similarly in terms of genuine significance; to pass from a ratio of 20 to one of 30 is greater than to pass from one of 30 to one of 40.

To illustrate certain of the tricky features to be guarded against in the use of ratios the following data and graphs are given:

CHART VII

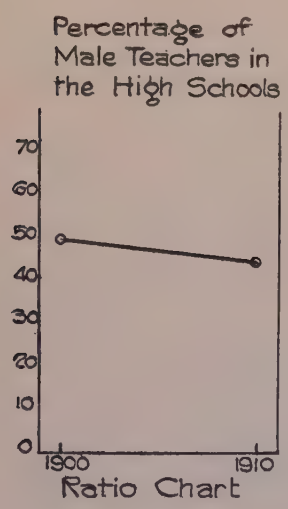


CHART VII a

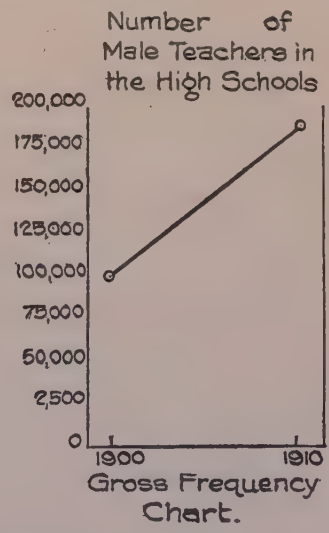


TABLE XIII

Number of Teachers in the Public High Schools of the U. S. Report of the Commissioner of Education, 1913, v. 2, pp. 9–10

	1900	*1910*
MEN . .	10,172 = 50 per cent of total 49.931 more exactly	18,890 = 45 per cent of total
WOMEN .	10,200 = 50 per cent of total	45.336 more exactly
TOTALS .	20,372	22,777 = 55 per cent of total
		41,667

From a casual glance at these charts it would be hard to realize that they are both accurate representations of the same data. A few pertinent questions might be asked:

(1) If the tendency shown by the ratio chart (tendency based upon the actual data for 1900 and 1910) continues, what will be the proportion of male teachers in the year 2000? Answer .03981.

(2) If the tendency shown by the gross frequency chart (tendency based upon the same actual data) continues, how many male high school teachers will there be in the year 2000? Answer 97,352.

(3) With the proportion as shown in your answer to question (1) and the number of male teachers as given in your answer to question (2), how many women teachers would there be in the high schools in the year 2000? Answer 2,348,064.

If the reader sees through this situation he appreciates one of the fallacies likely to arise through the use of proportions. Another occurs in combining ratios

Time Ratios

To average a number of ratios to obtain a single index, in general leads to an error. This will be considered later, but to illustrate the fact that ratios do not group themselves in a symmetrical manner around their own mean, the following data from Mitchell are given as quoted by Secrist. (1917, p. 312.) They also provided the material for an important problem in plotting.

It will be noticed that the class intervals extend over ranges of two units, e.g., there are five class intervals in covering a rise in prices from 10 per cent to (but not including) 20 per cent. With no direction to the contrary it is to be presumed that the class designated in the table by "54 – 55.9" includes all measures with values between the limits 53.95 and 55.95; that the next class includes measures between 51.95 and 53.95; etc. This is to say that presumably the data have been recorded to but one decimal place so that such measures as 53.86 and 53.92 are called 53.9 and a measure such as 53.96 is recorded as 54.0. If the recorder encountered a measure 53.95 he had to arbitrarily decide whether it would be called 53.9 or 54.0.

TABLE XIV

Distribution of 5578 Cases of Change in the Wholesale Prices of Commodities from One Year to the Next

Per Cent of Change from the Average Price of the Preceding Year (Falling Prices)	Number of Cases	Per Cent of Change from the Average Price of the Preceding Year (Rising Prices)	Number of Cases
54–55.9	1	14–15.9	106
—	—	16–17.9	102
50–51.9	1	18–19.9	73
48–49.9	1	20–21.9	65
46–47.9	1	22–23.9	45
44–45.9	2	24–25.9	47
42–43.9	4	26–27.9	29
40–41.9	5	28–29.9	30
38–39.9	5	30–31.9	22
36–37.9	7	32–33.9	17
34–35.9	10	34–35.9	18
32–33.9	7	36–37.9	11
30–31.9	16	38–39.9	17
28–29.9	27	40–41.9	14
26–27.9	17	42–43.9	6
24–25.9	32	44–45.9	10
22–23.9	39	46–47.9	11
20–21.9	45	48–49.9	5
18–19.9	71	50–51.9	1
16–17.9	76	52–53.9	4
14–15.9	107	54–55.9	3
12–13.9	120	56–57.9	1
10–11.9	173	58–59.9	6
8– 9.9	200	60–61.9	4
6– 7.9	238	—	—
4– 5.9	329	66–67.9	4
2– 3.9	375	68–69.9	3
Under 2	405	70–71.9	1
No change	697	72–73.9	4
(Rising Prices)		74–75.9	1
Under 2	410	—	—
2– 3.9	355	80–81.9	1
4– 5.9	356	82–83.9	1
6– 7.9	261	84–85.9	1
8– 9.9	237	86–87.9	1
10–11.9	167	—	—
12–13.9	115	100–101.9	1
		102–103.9	1
			5,578

For the data in hand it is not known how such a case would have been decided, but a very good rule to follow is to always assign such a critical measure to the even instead of the odd

value, i.e., the measures 53.95, 54.05, 54.15, 54.25, 54.35 and 54.45 would be assigned as 54.0, 54.0, 54.2, 54.2, 54.4 and 54.4 respectively. It will be noticed that in the long run this introduces no systematic error for the $\frac{1}{2}$ is thrown away as often as it is added. It does result in a slight piling up of the even measures, but that is generally inconsequential, whereas the adding of a half every few measures would result in a cumulative error which might be serious.

If the class intervals run in order from 53.95 to 55.95, 51.95 to 53.95, . . . 1.95 to 3.95 it is found that the next frequency, in order to extend over the same range, would be from − .05 to 1.95, i.e., from an increase in price of .05 per cent to a decrease of 1.95 per cent. This, however, cannot be the case, as a very large frequency, 697, is recorded for "no change." The way the data are recorded would suggest a class interval corresponding to "no change," but this cannot be so, as the intervals on either side preëmpt the space. In plotting the data, therefore, the "no change" interval must be squeezed out and its frequency, 697, distributed between the neighboring classes. We will assign 348 to the "under 2 — Falling prices" interval, and the remainder, 349, to the "under 2 — Rising prices" interval. There still is a slight discrepancy (.05) in the ranges of these two middle intervals, but as it cannot be positively accounted for without recourse to the original data it is passed over.

For convenience in tabulation and plotting we will consider the first class interval to extend from 54.00 to 56.00 and to have its mid-point or class symbol 55.00, the second a mid-point at 53.00, etc., and the frequencies as before.

The frequency polygon seems better suited to the data in hand, as it gives the impression of a more pronounced mode than would a histogram and in this case this feature should be emphasized.

Three ways of connecting the points of a distribution have been presented: (*a*) by drawing a histogram — Chart I; (*b*) by drawing a frequency polygon — Chart II; (*c*) by drawing a smooth curve through or near all the points which fits the data as nearly as can be determined visually — Chart III. A fourth way (*d*) is to plot from smoothed data; and a fifth (*e*) is by mathematically determining the equation of

the curve which best fits the data and plotting the same. This last method is discussed in Chapter VII. Methods (*a*), (*b*), and (*e*) preserve areas, i.e., the total area under the curve is equal to the population, or number of cases. Method (*e*) also preserves other important features. In using method (*c*) there should be a definite attempt to preserve areas; that is,

TABLE XV

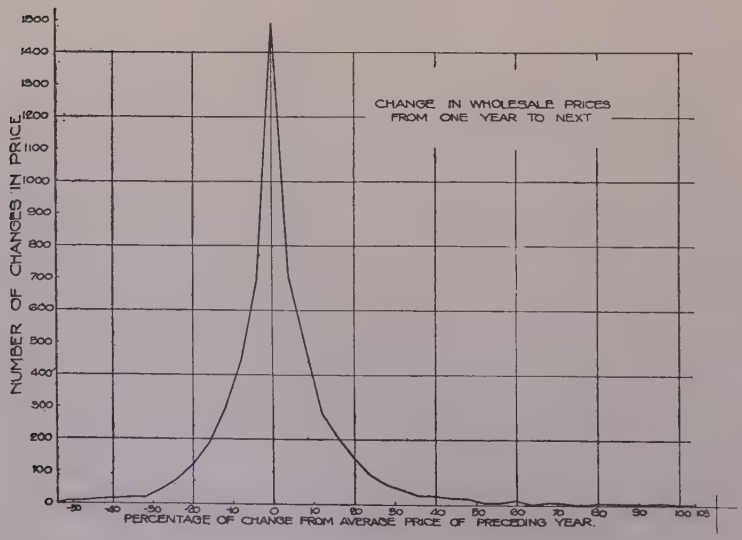
Per Cent of Change from the Average Price of the Preceding Year	Number of Cases	Class Interval of 4 Per Cent		Per Cent of Change from the Average Price of the Preceding Year	Number of Cases	Class Interval of 4 Per Cent	
		Per Cent of Change	Number of Cases			Per Cent of Change	Number of Cases
		− 56	1	− 5	329		
− 55	1					− 4	704
				− 3	375		
		− 52	1	− 1	753		
− 51	1					0	1512
− 49	1			1	759		
		− 48	2	3	355		
− 47	1					4	711
− 45	2			5	356		
		− 44	6	7	261		
− 43	4					8	498
− 41	5			9	237		
		− 40	10	11	167		
− 39	5					12	282
− 37	7			13	115		
		− 36	17	15	106		
− 35	10					16	208
− 33	7			17	102		
		− 32	23	19	73		
− 31	16					20	138
− 29	27			21	65		
		− 28	44	23	45		
− 27	17					24	92
− 25	32			25	47		
		− 24	71	27	29		
− 23	39					28	59
− 21	45			29	30		
		− 20	116	31	22		
− 19	71					32	39
− 17	76			33	17		
		− 16	183	35	18		
− 15	107					36	29
− 13	120			37	11		
		− 12	293	39	17		
− 11	173					40	31
− 9	200			41	14		
		− 8	438	43	6		
− 7	238					44	16

25987

TABLE XV (*continued*)

Per Cent of Change from the Average Price of the Preceding Year	Number of Cases	Class Interval of 4 Per Cent		Per Cent of Change from the Average Price of the Preceding Year	Number of Cases	Class Interval of 4 Per Cent	
		Per Cent of Change	Number of Cases			Per Cent of Change	Number of Cases
45	10			73	4		
47	11			75	1		
		48	16	—	—	76	1
49	5					80	1
51	1			81	1		
		52	5	83	1		
53	4					84	2
55	3			85	1		
		56	4	87	1		
57	1					88	1
59	6			—	—		
		60	10			92	0
61	4					96	0
—	—					100	1
		64	0	101	1		
67	4			103	1		
		68	7			104	1
69	3						
71	1				5,578		5,578
		72	5				

Chart VIII



if the curve as drawn lies above any point it should lie below some other, or, more accurately, the sum of the vertical distances which it lies above points in the actual distribution should equal the sum of the distances which it lies below other points. In drawing a free curve for incomes, Chart III, the preservation of total area is a difficult thing to insure, but for maximum temperatures, Chart I, it can be accomplished with fair accuracy and little trouble. The personal element which enters into method (*c*) generally makes it inadvisable for published work; but for original, hasty and personal research it may well be the one most frequently used.

Section 6. Smoothing Data

The smoothing of data preparatory to plotting (Method *c*) may be illustrated by the accompanying records of the U. S. Weather Bureau for New York City:

TABLE XVI

Mean Monthly Temperatures for 1917

Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
32.4	27.8	38.7	47.2	53.2	68.3	74.1	74.6	63.0	52.0	41.2	25.0

We have here a temporal series, and as is frequently the case, periodic fluctuations are shown. To obtain a general idea of variations within the year the curve at the end of December should join on to the curve at the beginning of January, as indicated below in Chart IX drawn by Method (*c*).

It will be noticed that in the 1917 data there is a minor mode in January and a major mode in August. As such bi-modality is not typical we will smooth by means of the moving average method and plot the resulting series. The moving average method consists of replacing original items by averages of a certain number of class frequencies. In the present problem we will average the frequencies for two neighboring class intervals and assign the result to the point midway between the two frequencies. If we consider the averages for each month as belonging to the 15th day of the month, we can take the average of the temperatures for January and February and assign this average to the end of January or the first of February. Next the February and March temperatures are averaged and the result assigned to March 1. Continuing

throughout the series, finally averaging the temperatures for December and January, gives the data of Table XVI, indicated on Chart IX by the ×'s.

CHART IX

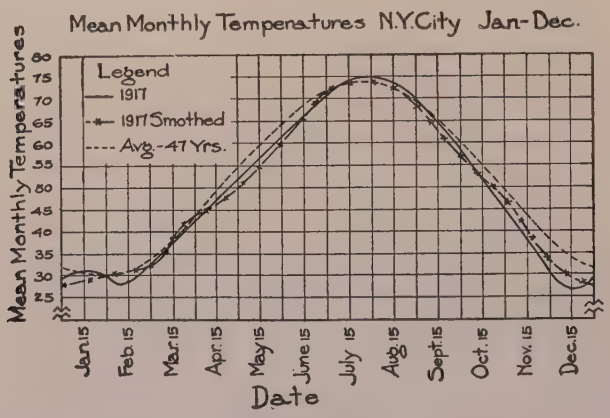


TABLE XVI–a

Mean Smoothed Temperatures for 1917

Ja. 1	F. 1	M. 1	A. 1	M. 1	J. 1	J. 1	A. 1	S. 1	O. 1	N. 1	D. 1
28.7	30.1	33.2	43.0	50.2	60.8	71.2	74.4	68.3	57.5	46.6	33.1

The reason this process is called that of taking a "moving average" would be better exemplified if groups of three or more items were averaged, in which case each successive sum is obtained from the preceding one by dropping one item and adding a second. It will be noticed that this curve has but a single mode, is much more regular than the curve from the original data, and does not have as high a maximum or as low a minimum, which fact is a necessary consequence of the method of smoothing. Moreover, it represents the annual fluctuations better than the curve from the original data, as is shown by comparing it with the dotted line based upon the records for the 47 years from 1871–1917, given herewith:

TABLE XVII

Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
31.0	30.5	37.8	48.7	59.8	68.8	74.0	72.6	66.4	55.7	43.9	34.0

Since the average of two unequal numbers is never as large as the larger or as small as the smaller of the two, the smoothing

process tends to flatten a curve out and lower modes. If the data are particularly irregular it is frequently desirable to do this to reveal a general trend, but it should be borne in mind that something of significance is always lost in the process of smoothing. Numerical calculations should never be made from

TABLE XVIII

Distribution of Marks given to Women in 8 Elective College Subjects. Below 60 Failure; 60–74 Condition

(Taken from Mary Theodora Whitley, A Statistical Study of College Marks — Master's Dissertation, Columbia, 1906)

Grade	*f* (Frequency)	*f* Av. of Three	*f* Av. of Five	*f* Av. of Fifteen	Grade	*f* (Frequency)	*f* Av. of Three	*f* Av. of Five	*f* Av. of Fifteen
43				.07	75	27	11.7	9.4	11.13
44				.06	76	7	15.0	12.4	13.47
45				.07	77	11	11.3	15.4	14.33
46				.06	78	16	14.3	21.0	18.00
47				.07	79	16	29.0	20.0	20.00
48			.2	.13	80	55	24.3	20.4	22.80
49		.3	.2	.14	81	2	23.4	24.6	26.07
50	1	.4	.2	.13	82	13	17.3	24.2	27.53
51		.3	.2	.14	83	37	21.3	26.8	32.07
52			.2	.13	84	14	39.7	32.6	32.00
53			.2	.53	85	68	37.7	38.6	34.73
54		.3	.2	.54	86	31	47.3	41.2	37.00
55	1	.4	.2	.53	87	43	41.3	43.0	37.07
56		.3	.2	.53	88	50	38.7	48.4	37.60
57			.2	.54	89	23	56.0	43.4	39.00
58			1.2	.93	90	95	41.3	45.2	38.87
59		2.0	1.2	.93	91	6	51.0	45.2	36.73
60	6	2.0	1.2	1.17	92	52	36.0	44.0	35.87
61		2.0	1.2	1.20	93	50	39.7	37.6	31.33
62			1.2	1.27	94	17	43.3	41.0	29.26
63			1.4	2.07	95	63	34.4	32.8	26.40
64		2.3	1.4	2.13	96	23	32.3	23.8	23.07
65	7	2.4	1.8	2.20	97	11	13.0	20.6	21.53
66		3.0	2.2	2.27	98	5	5.7	8.0	15.20
67	2	1.3	2.4	2.33	99	1	2.0	3.4	14.80
68	2	1.7	3.6	3.73	100		.3	1.2	11.33
69	1	5.3	3.8	4.20	101			.2	8.00
70	13	5.0	3.6	4.93	102				6.87
71	1	5.0	3.4	6.00	103				2.67
72	1	1.0	3.4	7.07	104				1.13
73	1	1.0	6.2	10.27	105				.40
74	1	9.7	7.4	10.40	106				.07
						773.	773.	773.	773.

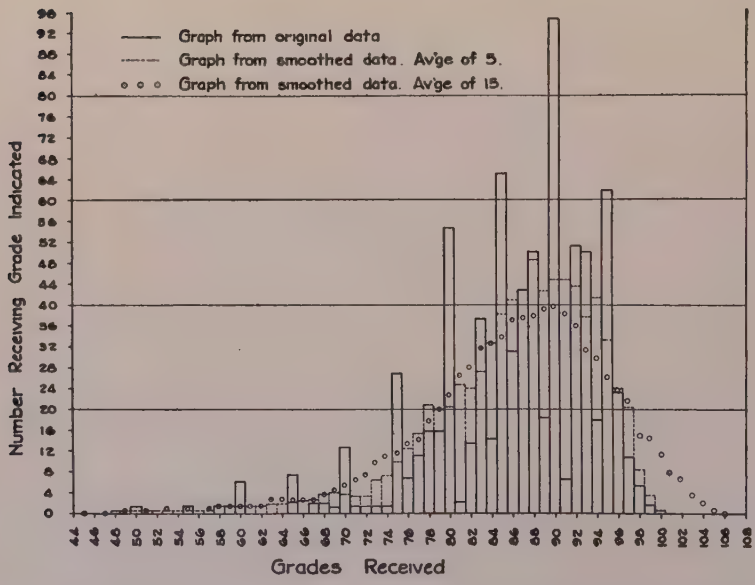
smoothed data, as a spurious consistency in the findings may be introduced and significance of the original data may be hidden.

The possibilities and limitations of smoothing will be better illustrated by application to the data of Table XVIII which are decidedly multi-modal.

In the accompanying Chart X, the histogram represents the original data; the smoothed average-of-three curve is not

CHART X

Distribution of School Grades


Graph from original data
Graph from smoothed data. Av'ge of 5.
Graph from smoothed data. Av'ge of 15.
Number Receiving Grade Indicated
Grades Received

shown; the ordinates of the smoothed average-of-five curve are represented by dots; and the ordinates of the smoothed average-of-fifteen curve are represented by o's.

The curve from the original data has fourteen modes, ten of them located at grades divisible by five and four located halfway between such grades. It seems that many teachers do not grade on a percentile scale in units smaller than five per cent, and that most of the remainder do not grade in units less than two and one half per cent. An examination of the frequencies in the average-of-three column shows that these minor modes, which occurred about every $2\frac{1}{2}$ units, have been

smoothed out by the process of averaging three neighboring measures, but that all the major modes persist though they occasionally are no longer exactly five units apart. It is found, by reference to the plotted distributions, that it requires the smoothed average-of-five curve (· · · ·) to smooth out the modes periodically occurring every five units. It is also apparent that the smoothed average-of-fifteen curve has flattened the mode at 90 and spread out the extreme measures altogether too much. It is therefore a desirable rule, when smoothing must be resorted to, to average such a number of neighboring groups as just cover the periodicity which it is desired to smooth out. If the data show great irregularity, rather than periodicity, it is better to average too small a number of groups than too large a number. In the case in hand there is no doubt that the smoothing by averaging five class frequencies is the preferable method, but even so, something of significance, as is always the case, has been lost by the smoothing: To illustrate; the percentage of failures shown by the smoothed data, .57 per cent, is over twice as large as was in reality the case, — .26 per cent.

Section 7. The Ogive Curve

When it is desired to determine the number of cases or per cent of the population lying below a certain record, it can be readily done if a curve is plotted showing sums of the frequencies of all measures below designated amounts of the trait. The method may be illustrated by the data of Table I. The first two columns below repeat that table; the third column is obtained by cumulating the frequencies in column two. The 1 in column three recorded opposite 65.5 means that one day (out of the 62) had a temperature less than 65.5. It will be noticed that two days had temperatures less than 66.5, or 67.5, or 68.5, or 69.5. In such a case it is sounder to assign the 2 to the point midway between the 65.5 and the 69.5 than to any other point in this stretch. Accordingly it is recorded in column three that 2 days had temperatures less than 68.0. Continuing there are 3 days with temperatures less than 70.5; 4 with less than 72.5, etc. Finally it is to be noted that the last point is indeterminate, i.e., 62 days had temperatures

less than 98.5, or 99.5, or 100.5, etc. It is impossible to determine from finite data what is the maximum temperature below which the temperatures for all days lie. It is of course also impossible to determine what is the minimum temperature above which the temperatures for all days lie. For this

TABLE XIX

Distribution of Daily Maximum Temperatures, July and August, New York City, 1917

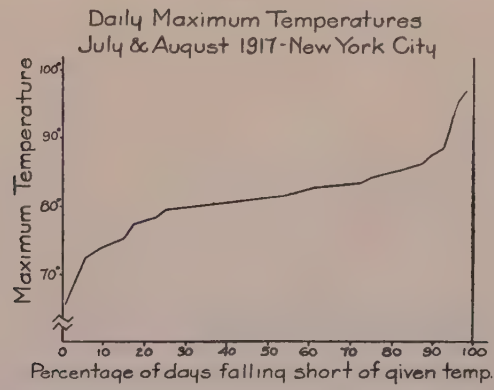
Temperatures	No. of Days with Given Temperature	Cumulations of No. of Days	Cumulations Expressed in Percentages	Temperatures	No. of Days with Given Temperature	Cumulations of No. of Days	Cumulations Expressed in Percentages
65	1					45	72.6
65.5		1	1.6	84	2		
66	1					47	75.8
66.5				85	4		
67						51	82.3
68		2	3.2	86	3		
69						54	87.1
70	1			87	1		
		3	4.8			55	88.7
71	1			88	2		
72							
		4	6.4	89		57	91.9
73							
74	2			90	1		
		6	9.7	91			
75	3			92			
		9	14.5			58	93.6
76	1			93			
		10	16.1	94			
77	1			95	1		
		11	17.7			59	95.2
78	3			96	1		
		14	22.6	97		60	96.8
79	1			98	2		
		15	24.2			62?	
80	10			99			
		25	40.3			62?	
81	8			100			
		33	53.2			62?	
82	5			101			
		38	61.3				
83	7						

reason the zero and one hundred percentile points for this ogive curve are not plotted. This should be the case for all ogive curves — the common practice of plotting the lowest

and highest recorded data as the 0 and 100 percentiles being inaccurate and confusing.

Column four gives the same data as column three, expressed in percentages of the total frequency. In the accompanying graph the ordinates are the cumulative frequencies in percentages and the abscissas are the temperatures as shown:

CHART XI


Daily Maximum Temperatures
July & August 1917 - New York City
Maximum Temperature
100°
90°
80°
70°
0 10 20 30 40 50 60 70 80 90 100
Percentage of days falling short of given temp.

It is interesting to note that the relatively irregular data used has resulted in a fairly regular ogive curve, and that, without any smoothing. The ogive curve facilitates interpretation, e.g., it is immediately read from the curve that:

5 per cent of the days do not attain a temperature of 71°
10 " " " " " " " " " " " 75°
20 " " " " " " " " " " " 78°
50 " " " " " " " " " " " 81°
90 " " " " " " " " " " " 88°
95 " " " " " " " " " " " 95°
50 " " " " " have maximum temperatures between 79.5° and 84.5°, etc., etc.

Or, interpolating the other way:

A temperature of 95 or more is reached on 5 per cent of the days
" " " 85 " " " " " 21 " " " " " "
" " " 75 " " " " " 87½ " " " " " etc.

The ogive curve may also be used to determine the mode, for if a smooth curve (not a polygon as here shown) is drawn through or near the points given and a ruler rotated so as to be tangent to the curve at successive points, that point at

which the ruler ceases turning in one direction and starts to turn in the other (called the point of inflection) is the modal point, its value being read from the ordinate measures on the margin. Applying this method to these particular data the mode is found to be very close to 81°. The more important measures revealed by the curve are the median, or 50-percentile, the semi-interquartile range more briefly called the quartile deviation, or one half the distance between the upper and lower quartiles, the 10-percentile, the 90-percentile and the 10–90-percentile range. For the data in hand these are respectively 81°, 79.5°, 84.5°, 2.5°, 75°, 88° and 13°.

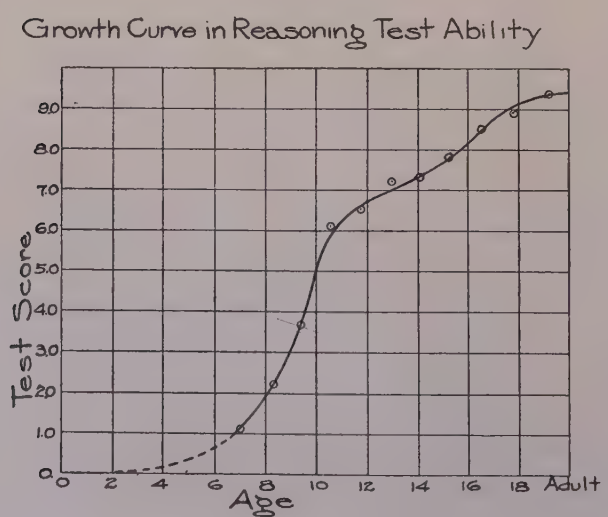
Section 8. The Growth Curve

The accompanying table gives smoothed scores in a reasoning test as given by Kelley (1917). Plotted they give a typical growth curve.

TABLE XX

Age .	7.0	8.3	9.4	10.5	11.8	13.0	14.1	15.3	16.5	17.8	Adult 19.2
Score on Trabue Scale .	1.1	2.2	3.7	6.1	6.5	7.2	7.3	7.8	8.5	8.9	9.4

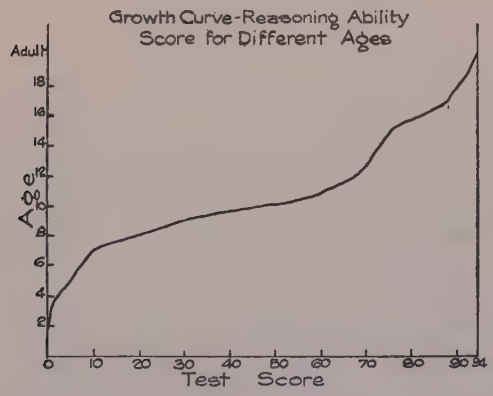
Chart XII



This particular curve is interesting in that it snows a flattening at ages 13 and 14, which is not at all characteristic of growth curves of mental traits, but as the units of measurement, instead of intrinsic ability, could conceivably account for the phenomena the curve does not prove, but merely suggests, that there is a pubertal disturbance. For the purpose of the present statistical treatment no attention need be paid to the double inflection of the curve.

Rotating the curve through 90° and looking at it in a mirror (as pictured in Chart XIII) shows its general resemblance to an ogive curve. It was possible in the case of daily temperatures to cumulate scores and obtain ogive curve data. By the reverse process it is possible from the ogive data to obtain the original distribution of temperatures. By parity of operation it is possible to obtain measures of growth increments from an original growth curve. The growth curve may be plotted as herewith:

CHART XIII



Thinking of the abscissas as sums of increments of reasoning ability and recalling that the graph is for an average individual, whose maximum development or accumulation is to 94 of such increments (i.e., the total population of increments is 94) the graph may be read: At age 7 the individual possesses 11 increments of reasoning ability; at age 10, 50 increments,

etc. This may be an awkward way of interpreting growth, but if it is desired to think of growth as a sum of increments it immediately suggests the determination of the increments added during each year of life as follows:

TABLE XXI

Age	Score	Yearly Growth Increment	Age
0	0		
		0	.5 (from 0–1)
1	0		
		.5 −	1.5 (from 1–2)
		.5 +	2.5 (from 2–3)
3	1		
		2 −	3.5 (from 3–4)
		2 +	4.5 (from 4–5)
5	5		
		3 −	5.5 (from 5–6)
		3 +	6.5 (from 6–7)
7	11		
		8	7.5 (from 7–8)
8	19		
		12	8.5, etc.
9	31		
		19	9.5
10	50		
		12	10.5
11	62		
		5	11.5
12	67		
		4	12.5
13	71		
		2	13.5
14	73		
		3	14.5
15	76		
		6	15.5
16	82		
		5	16.5
17	87		
		3	17.5
18	90		
		4	? (from 18–adulthood)
Adult	94		
		—	
		94	

These growth increments plotted in the form of an ordinary frequency polygon give the following figure:

CHART XIV

Distribution of Yearly Growth Increments in a Reasoning Test

The bi-modality of the growth increment curve is of course a consequence of the double inflection of the growth curve. Since the constants of this increment curve (mean, skewness, standard deviation, etc.) can be readily calculated, the curve has certain advantages over the growth curve. It should be a very convenient form in which to present data for purposes of studying variability in rate of growth, variability in price changes, etc. In dealing with functions in which there is a loss in a given period, e.g., when an individual weighs less in one year than in the preceding, negative frequencies arise. These need cause no trouble if treated strictly algebraically and the negative sign preserved.

Brown and Thomson (1921) have shown that the standard deviations of the class frequencies of such a curve are not given by the ordinary formula [Formula 25].

Section 9. The Graphic Representation of Categorical Measures

The graphs thus far have pictured the frequencies or amounts of a quantitative or temporal variable, but if the frequencies of categorical measures are desired a different procedure is necessary. For example, if desired to represent the number of

days which lie in the following categories, (*a*) clear, (*b*) cloudy, (*c*) rainy, a large number of devices are possible. The significant feature to be portrayed in this as in all qualitative series is the magnitude of each category with reference to the others, or the proportions which each bears to the whole. This may be shown by appropriate lengths of lines constituting what is called a "bar diagram," by heights of shaded rectangles, by sectors of the required number of degrees, by appropriate number of discrete objects, men, bushels, ships, etc. The essence of an accurate portrayal lies in having the representations of the two or more items alike in every respect except one and differing in that one by the required amounts.

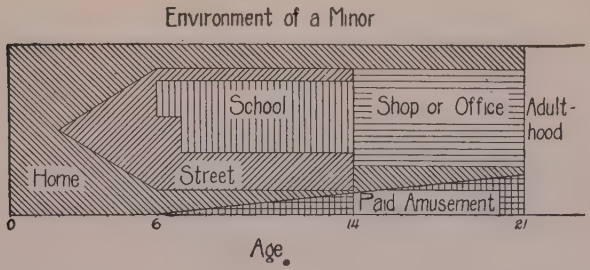
If the population of Texas is 5 million and that of Georgia 3 million and if a man, representing Texas, is pictured beside a child, three fifths as tall, representing Georgia, the impression conveyed is entirely erroneous. The heights are in the ratio of 5:3, but the areas covered by the figures are approximately in the ratio of 25:9. However the situation is even worse than this for the weight of a man as pictured is to the weight of a child as pictured approximately as 125:27 and one is inclined, in so far as the pictures mean a man and a child, to make just such a comparison.

If three dimensional objects are pictured upon a two dimensional surface to convey a one dimensional relation the objects should be identical in size and differ only in number. In the illustration mentioned, Texas could be represented by a row of five men and Georgia by a row of three. The use of men in picturing population, of sectors of a dollar in showing the items of a budget, of bales of cotton in picturing cotton production, etc., are conventional and expressive modes of presentation. Accuracy of presentation is favored by the use of rectangles of different lengths, but as independence of a heading may be accomplished by a proper choice of object for picturization, this method has certain indubitable advantages. However, if a two or three dimensional object is pictured either (*a*) all the dimensions except one should be kept constant and that one vary in the proportions desired, or (*b*) all dimensions should be the same and the number of objects vary. As an illustration of (*a*) the amount of paving in two cities could be repre-

sented by the pictured lengths of two roads, the amount of coal produced by trains of gondola cars of different lengths, or the number of fish in the lakes at two resorts by angle worms of different lengths, etc.

It is occasionally possible to represent not only the relative size of two categories but also their special temporal or spatial relation by graphic means. This is very prettily illustrated by the accompanying figure from Perry. (C. A. Perry, Educational Extension. Quoted by Rugg, 1917.)

CHART XV



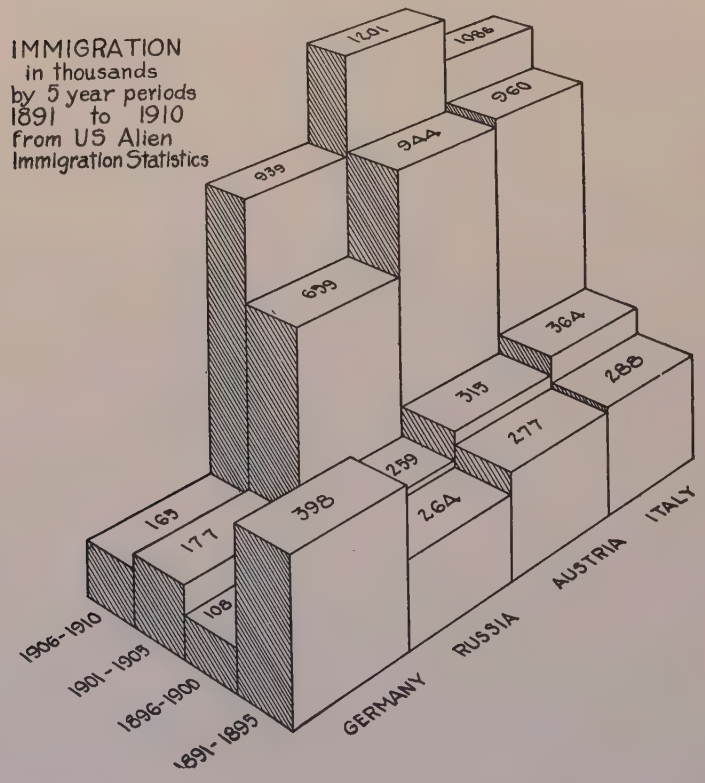
A cross section at any age reveals the proportions of time spent in the various ways, but it does more than this, as it reveals the temporal relations of these proportions.

If one considers how many pages of writing matter would be required to convey an idea of all the relationships shown in Chart XV he will appreciate the art involved in graphic presentation. If he will likewise consider that a written presentation would probably be obscure and dreary reading and that the joy of discovery belongs to one who studies an ingenious chart, he will appreciate that the graphic method at its best has far greater advantages than those of simply saving space and time.

The last figure conveyed information as to three different items, (*a*) age, (*b*) time spent in different activities, and (*c*) the temporal disposition with reference to each other of different activities. It is thus a complex series, being quantitative, qualitative and temporal.

Accompanying is a block presentation of a complex series. It conveys information as to three different things, (*a*) date, (*b*) numbers of immigrants, and (*c*) country of birth.

CHART XVI

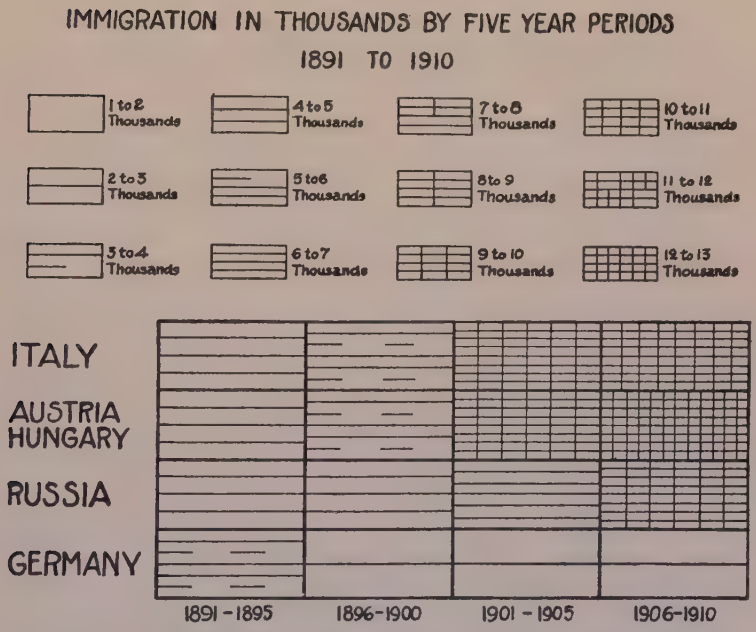


This information is fully presented in the figure, but it very frequently is impossible clearly to present a three-dimensional situation by a picturization of a three-dimensional figure, for commonly a part of the figure would obscure other essential parts. The large immigration from Germany in 1891–95 almost hides the block showing the immigration from Germany in 1896–1900, but as it does not completely hide it the relationships are readily apprehended. However, if the immigration

from Russia had also been larger in 1891–95 than in 1896–1900, the block for the latter period would not have been visible and the method would have been unsatisfactory.

Another device for presenting such data is given below:

CHART XVII



This is a more flexible method than the preceding, as there is no possibility of one block covering up another, but it requires a coarse grouping in the measure represented by the cross-hatchings, or shadings, and in general its features are not outstanding as are those of the preceding figure.

In the block figure the last three countries are in the order demanded by geographical position of the countries. An additional fact, such as the numbers of literate and illiterate immigrants, could be represented by shadings of appropriate areas upon the tops of the blocks. Still another, such as age, or sex, or vocation, could be shown by the color of the ink used in the cross-hatching. Even this does not exhaust the possibilities of graphic presentation upon a single two-dimensional surface.

It is difficult to give a summary of the principles underlying graphic portrayal as they differ with the number of dimensions presented and with the continuous or discrete nature of the data, but the recommendations contained in the preliminary report of the joint committee on standards of graphic presentation are of broad applicability. This committee represented a wide field of statistical workers and was formed upon the invitation of the American Society of Mechanical Engineers. Its recommendations as given by Haskell (1919) are:

1. The general arrangement of a diagram should proceed from left to right.

2. Where possible represent quantities by linear magnitude, as areas or volumes are more likely to be misinterpreted.

3. For a curve the vertical scale, whenever practicable, should be so selected that the zero line will appear in the diagram.

4. If the zero line of the vertical scale will not normally appear in the curve diagram, the zero line should be shown by the use of a horizontal break in the diagram.

5. The zero lines of the scales for a curve should be sharply distinguished from the other coördinate lines.

6. For curves having a scale representing percentages, it is usually desirable to emphasize in some distinctive way the 100% line or other line used as a basis of comparison.

7. When the scale of the diagram refers to dates, and the period represented is not a complete unit, it is better not to emphasize the first and last ordinates, since such a diagram does not represent the beginning and end of time.

8. When curves are drawn on logarithmic coördinates, the limiting lines of the diagram should each be at some power of 10 on the logarithmic scale.

9. It is advisable not to show any more coördinate lines than necessary to guide the eye in reading the diagram.

10. The curve lines of a diagram should be sharply distinguished from the ruling.

11. In curves representing a series of observations, it is advisable, whenever possible, to indicate clearly on the diagram all points representing the separate observations.

12. The horizontal scale for curves should usually read from left to right and the vertical scale from bottom to top.

13. Figures for the scale of a diagram should be placed at the left and at the bottom or along the respective axes.

14. It is often desirable to include in the diagram the numerical data or formulæ represented.

15. If numerical data are not included in the diagram it is desirable to give the data in tabular form accompanying the diagram.

16. All lettering and all figures in a diagram should be placed so as to be easily read from the base as the bottom, or from the right-hand edge of the diagram as the bottom.

17. The title of a diagram should be made as clear and complete as possible. Sub-titles or descriptions should be added if necessary to insure clearness.

PROBLEMS

1. Smooth the temperature data by means of a moving average of three class frequencies and plot. What is the modal value?

2. Express the populations of California, Oregon and Washington as indexes with 1900 as base. Which state showed the greatest relative growth in the decade 1900–1910?

3. Chart VI shows that relative to 1907 retail prices of steak in Chicago did not advance as fast as wholesale prices. Choosing each year in turn as base, determine the relative increase in the wholesale prices and the retail prices of steak for the succeeding year, and answer the question, "In how many years did retail price advances fail to keep pace with wholesale price advances?" Using data in the last column of Table XII a answer the same question with reference to Wholesale prices and Retail prices of 22 common articles.

4. (*a*) Plot an Ogive curve for the raw data of Table XVIII and on the same paper (*b*) an Ogive curve for the same data as smoothed by a moving average of fifteen class frequencies.

5. Plot hypothetical data giving incomes in Great Britain in the form of an Ogive curve. What is the mode? Fill out the following table:

Incomes Received by Successive Percentiles

Percentiles	1	5	10	20	25	30	40	50	60	75	80	90	95	99
Incomes														

6. Save work for future reference.

CHAPTER III

THE MEASUREMENT OF CENTRAL TENDENCIES

Section 10. Averages

A tabulation of the data pertaining to a distribution presents all the facts, and a histogram or frequency polygon makes possible the visualization of this detail. Ordinarily, however, the detail is so great that it cannot be interpreted. In this case certain measures of the total distribution are serviceable in summarizing the data. The most important of these are averages, or measures of central tendency. The most significant averages are (*a*) the mean [more accurately the arithmetic mean], (*b*) the median, (*c*) the mode, (*d*) the geometric mean, and (*e*) the harmonic mean. Note that these are all averages. The word "average" is frequently used synonymously with mean (arithmetic mean). It will occasionally be used in this text in such expressions as "the average of the means," in order to avoid the more accurate but awkward expression "the mean of the means." Ordinarily "mean" will be used consistently to designate the arithmetic mean, and "average" as synonymous with "measure of central tendency," thus meaning any one of the five measures listed above.

The most important single item of information to be known about a distribution is what it is a distribution of.

The second in importance is the number of cases in the distribution, or, as it is usually expressed, the population.

The third, is to know some measure of central tendency, some average.

The fourth, to know some measure of the degree to which the measures scatter, or lie above and below the average, i.e. to know a measure of dispersion or deviation from the average.

The fifth, to know if the measures are symmetrically distributed with reference to the average, or if there is a bunching of measures on one side of the average and a long tailing out of measures on the other side; i.e., to know a measure of skewness.

The sixth, to know if the measures are exceptionally densely grouped at the average, giving a high peak to the frequency polygon (leptokurtic, i.e., β_2 of section 36 is greater than 3.0) or if the distribution is rather flat in the middle and contracted at the ends, thus tending toward a rectangular shape; (platykurtic, i.e., $\beta_2 < 3.0$) or if they show a mean between those two conditions as does a normal distribution (mesokurtic, $\beta_2 = 3.0$); in short, to know a measure of kurtosis.

These are all of the essential measures in the case of a uni-modal distribution; the next important item would be a measure of the tendency to have more than a single mode, or place of dense frequency.

No treatment will be given in succeeding chapters of bi-modal curves, but if it is noted that uni-modal curves include anti-modal or U-shaped curves, — those having large frequencies at the extremes and small frequencies in the middle, as well as L-shaped curves, rectangular distributions, and all forms of positive uni-modal curves, it will be seen that the great majority of distributions found in biology, economics, and psychology belong to the uni-modal type and that a knowledge of the six items mentioned above is adequate for all but a small number of distributions.

Measures of skewness and kurtosis are essential in mathematically fitting curves to observations and are treated of in Chapter VIII on Curve Fitting. The calculation of averages is dealt with in this chapter and the relative excellence of the different averages will be considered in connection with their probable errors in the next chapter.

Section 11. The Arithmetic Mean

The mean may be defined as the sum of the separate measures divided by the number of them. This definition immediately suggests the method of calculation: add the measures and divide by the population. If an adding machine is available and other measures of the distribution are not desired, this method is the most expeditious one to follow. Generally, however, it is more economical of time first to group the measures and arrange them according to magnitude, as was done with the Temperature data, Table VIII. Repeating

these data we have the first two columns of the accompanying Table XXII. The third column illustrates one method of

TABLE XXII

Calculation of the Mean

Temperatures	Frequencies	Products	Deviations from Arbitrary Origin	Products	Grouped Frequencies	Deviations from Arbitrary Origin	
X	f	fX	ξ^*	$f\xi$	F	ζ^*	$F\zeta$
65	1	65	− 15	− 15			
66	1	66	− 14	− 14	2	− 5	− 10
69					1	− 4	− 4
70	1	70	− 10	− 10			
71	1	71	− 9	− 9			
72					1	− 3	− 3
74	2	148	− 6	− 12			
75	3	225	− 5	− 15	6	− 2	− 12
76	1	76	− 4	− 4			
77	1	77	− 3	− 3			
78	3	234	− 2	− 6	5	− 1	− 5
79	1	79	− 1	− 1			
				− 89			− 34
80	10	800	0				
81	8	648	1	8	23	0	
82	5	410	2	10			
83	7	581	3	21			
84	2	168	4	8	13	1	13
85	4	340	5	20			
86	3	258	6	18			
87	1	87	7	7	6	2	12
88	2	176	8	16			
90	1	90	10	10	1	3	3
95	1	95	15	15			
96	1	96	16	16	2	5	10
98	2	196	18	36			
99					2	6	12
	62	5,056		185	62		50
		$M = 81.548$		− 89			− 34
				96			16
				62			62
			Correction = 1.548			Correction	
			Arbitrary			= 3 × .258	
			Origin = 80.			Correction = .774	
			$M = 81.548$			Arbitrary	
						Origin = 81.00	
						$M = 81.774$	

* Greek alphabet given in appendix.

calculating the mean; the fourth and fifth columns a briefer method, in that it involves handling smaller numerical magnitudes; and the last three columns another method which is still shorter in case the number of class intervals is large. For a method of calculating the mean, standard deviation, and higher moments by means of continued summations see Brown and Thomson (1921) and Elderton (1905).

The headings of these columns are typical and will be repeatedly used in subsequent examples:

X (or Y) will be used regularly, as here, to designate gross scores.

f (F) designates class frequencies.

ξ* (ζ) designates deviations of scores from an arbitrary origin, or starting point. In column four, ξ represents deviations of the gross scores from the arbitrary origin 80, while in column seven ζ represents deviations of class intervals each of which is three times as large as the class interval obtaining in the gross scores, e.g., from 0–1 in column seven is one ζ unit but it is three X units.

x (or y) has not been used in any of the above columns since it is reserved for a very definite purpose. It will consistently mean a deviation from the true mean. In the case in hand, if deviations from 81.548 had been recorded they would have been designated as x measures. Throughout the rest of this text x (or y) will mean a deviation from the mean or from an origin so near to the mean that no attention need be paid to the fact that it differs slightly from the true mean.

N. One further symbol is universally employed — N (n) stands for the population. In the present example $N = 62$. [n occasionally has other meanings, particularly when it appears as a subscript or a superscript.]

M is used to designate the mean.

Σ. The symbol Σ indicates not a measure but an operation. When placed before a symbol standing for a measure it indicates that the sum of all such measures is to be obtained, e.g., Σf means the sum of the frequencies — in the illustration $\Sigma f = 62$.

With these definitions in mind it will be seen that the mean

* Greek alphabet given in appendix.

may be calculated according to any one of the following formulas:

$$M = \frac{\Sigma X}{N}. \qquad \text{(The mean)}.....[1]$$

This formula is used in case measures are not grouped or arranged according to magnitude.

$$M = \frac{\Sigma f X}{\Sigma f}. \qquad \text{(The mean)}....[1\ a]$$

This is the method used in columns two and three. $\Sigma fX = 5056$ and $\Sigma f = 62$. These two formulas are really identical, for ΣfX simply means that each X is taken as many times as it occurs. There is no mathematical operation in use in which the sum of the measures is taken irrespective of the frequencies in the various classes, so that in subsequent examples ΣX will mean identically the same thing as ΣfX and will frequently be written for the latter as it is more concise. For similar reasons $\Sigma\xi$ will be written for $\Sigma f\xi$; Σx for Σfx; Σx^2 for Σfx^2; etc.

$$M = \text{Arbit. Orig.} + \frac{\Sigma \xi}{N}. \qquad \text{(The mean)}.....[1\ b]$$

This is the method illustrated in columns four and five. It is called the method of moments, i.e., of tendencies to produce rotation about a point. Moments may be taken about any origin and if the positive exceed the negative it means that the origin chosen is too small. Similarly if the negative exceed the position moments the guessed mean, or arbitrary origin, is too large and a negative correction is necessary. If the guessed mean is 80 and calculation shows that there are 96 excess positive moments then, since there are 62 cases in all, the moment corresponding to each measure should be $96/62 = 1.548$ greater than it is in order to make the positive exactly equal the negative moments. This point where the moments exactly balance is the mean. Obviously if the guessed origin is moved by 1.548 units, i.e., if 1.548 be added to 80, a value will be determined such that if moments about it are taken the negative and positive moments will exactly balance.

$$M = \text{Arbit. Orig.} + \frac{(\text{Class interval})\ \Sigma\ \xi}{N}. \qquad \text{(The mean)}\ [1\ c]$$

This method is illustrated in the last three columns. It is a moment method applied to data which have been grouped. The guessed origin is here 81, the class interval 3, i.e., 3 of the

gross measure units, and $\Sigma\zeta = 16$. Solving $M = 81.774$. The discrepancy between this value and that obtained before is due to the grouping, — the true value being 81.548 and not 81.774. Such error may be either positive or negative, and, unless very great precision is demanded, may be disregarded when the data show no pronounced periodic disturbances and when the number of class intervals is 12, or greater. (For considerations leading to the number 12 see section 46.)

It will be noted that there are 11 class intervals in column ζ. In the case of distributions which show peculiar local groupings great care should be exercised in combining class frequencies. In the case of the College Marks given in Table XVIII a combining of measures into groups as follows: 50.0 – 54.9, 55.0 – 59.9, 60.0 – 64.9, etc., and a designating of the middle points of the groups as lying at 52.5, 57.5, 62.5, etc., would lead to substantial error in calculating the mean, since the measures in the groups are not all evenly distributed. To illustrate: if the 12 measures in the interval 65.0 – 69.9 are grouped and assigned the value 67.5 an error of 1.33 has been introduced, for calculation shows that the true mean of these 12 measures is 66.17. An error of 1.33 in a single group would not be serious, but for the College Marks data the error is typical of each group, so that a calculation of the mean from data so grouped would lead to systematic raising of the mean by an amount between 1 and 2 units. Whenever systematic local tendencies are apparent in data and grouping is resorted to, it should be endeavored to so group that the middle of each group interval corresponds to a local mode; e.g., with the College Marks the class intervals of the groups should be as follows: 47.5–52.5, 52.5–57.5, etc., since the mid-points of these intervals, 50, 55, etc., correspond to local modes and also approximately to the means of the measures in the group intervals.

The data in Table XXIII reported by the New York State Industrial Commission and taken from the New York World of Jan. 27, 1919, are so grouped as to make it impossible accurately to determine any sort of an average wage. These data show that 6 per cent of women factory workers receive from \$6–\$7.99 a week, but certainly the mean wage of this group is not \$7.00, for in all probability a large number re-

ceive exactly $6.00, another large group exactly $7.00. while lesser groups receive wages of $6.50 and $7.50, and but very occasionally would there be a wage such as $6.49 or $7.99. Since one end of the interval, $6.00, has a large frequency that is not balanced by the other end $7.99, the mean of the group may be expected to lie below $7.00, possibly considerably below. Similarly the 14 per cent receiving wages from $8.00 to $9.99 presumably receive a mean wage much below $9.00. It is difficult to group data of this kind without introducing large error, but if the intervals had run, $6.25–$6.75, $6.76–$7.24, $7.25–$7.75, etc., probably the mid-point of each group would be close to the mean of the group. An attempt to determine an average wage from the data as given might easily be nearly 50 cents in error. The unequal distances covered by successive intervals in the grouping proposed is a disadvantage which is more than compensated by having the mid-points and the means of the groups approximately coincide.

TABLE XXIII

Full-time Earnings of 20,597 Women in Factories and 23,203 in Mercantile Establishments

	Factories Per Cent	Stores Per Cent
Less than $ 6	1	1
Less than 8	7	7
Less than 10	21	23
Less than 12	42	44
Less than 14	59	64
$14 or over	41	36
$20 or over	11	9

In any research the question usually arises whether to group at all, and, if so, what groupings to make. It has already been suggested that groupings should not be made which result in less than twelve classes. This is a lower limit. If the distribution is pronouncedly asymmetrical, as for example is that showing incomes in Great Britain, twelve is far too small a number of classes. The lower end of that curve could not be at all satisfactorily represented if the income range covered by each interval should be as large as £100, nor with such grouping could the arithmetic mean be accurately determined.

A range of £40 for the lower intervals will answer, though a range of £10 or £20 would be much better. Since incomes range from about £0 to £200,000 there would be no less than 5000 classes needed to represent the distribution if the class interval is £40.

The distribution of Wholesale Price Indexes is not as markedly asymmetrical, but it has such a phenomenal peak at "no change" that a coarse grouping cannot be used, or this characteristic is hidden. The plotted distribution has 41 classes and 38 of them have frequencies other than zero. As plotted, the peak at "no change" is less pronounced than it is in reality and if the grouping were coarser it would be still less apparent. A slightly coarser grouping would not have very great effect upon the mean, but it would have decided effect upon other constants, particularly those measuring kurtosis. Forty classes is close to the minimum which would be satisfactory for either graphic or numerical work with wholesale price index measures.

For graphic presentation of College Marks a grouping into classes of five units each, with interval limits chosen as already indicated, would result in a graph nearly as satisfactory as that based upon the moving average involving five neighboring classes. Such a grouping leads to but 11 classes, which is too small for very reliable results. However, groupings into units of 4, 3, or 2 are not satisfactory, as they do not conform to the local periodicity, which is five units. A grouping into units of $2\frac{1}{2}$ would be excellent from the standpoint of statistical accuracy, but as it would involve splitting the frequencies in the gross score classes it would be uneconomical of time. All things considered it would seem advisable to use the gross score intervals, or, for rough work, a grouping of five gross score intervals.

The situations presented by Incomes, Price Indexes, and College Marks are not typical, but illustrative of the more difficult grouping problems encountered.

Consider the Temperature data, Table XXII, and note that if two gross score intervals had been grouped the frequencies in Column *F* would have been for intervals whose mid-points would be 65.5, 67.5, 69.5, etc., that when three gross score

intervals are combined the mid-points are, as shown, 66, 69, 72, etc.; and that in general if an even number of gross score intervals are combined the mid-points of the resulting intervals do not coincide with the mid-points of any of the original intervals but lie halfway between original measures. Accordingly if an even number of gross score intervals are combined an entirely new table has to be made out. As this involves work and an additional chance for error it is undesirable if a grouping of an odd number of intervals will suffice.

As a general rule, applying to distributions not especially asymmetrical (skew) nor peaked (leptokurtric), (1) an odd number of gross score classes should be grouped, (2) the number of classes resulting from grouping should not be less than 12, and (3) the number of gross score intervals in a group should equal the number involved in local periods, or divide into such number without remainder, or be an integral multiple of such number. Finally in case the distribution is markedly skew or leptokurtic, conditions (1) and (3) remain the same but (2) the number of classes should be greater than 12 and great enough that significant portions of the distribution are revealed in such detail as is commensurate with their importance.

In determining the number of gross score intervals to be grouped in ordinary data a serviceable rule to follow is to subtract the smallest from the largest measure and divide by twelve. The nearest odd integer *below* the resulting quotient is the proper number of gross score intervals to combine. E.g., in the case of maximum temperatures $(98 - 65)/12 = 2.75$. The nearest odd integer below 2.75 is 1. Accordingly the data are not grouped at all and the gross score intervals of 1° kept as the proper steps. No material inaccuracy would have been introduced by combining two of the gross score intervals, but it would have been of questionable economy to do so.

Applying the rule to the College Marks data we have, $(99 - 50)/12 = 4.1$. The nearest odd integer below is 3. It would therefore be appropriate to group three intervals were it not for the fact that there is a local periodicity extending over 5 gross score intervals. Applying to wholesale price indexes $[103 - (-55)]/12 = 13.2$. Since the original scores were recorded in 2 per cent steps the interval of 13.2 per cent is

equivalent to 6.6 of the gross score intervals. The nearest odd integer below 6.6 is 5, which would be the proper number of gross score intervals to combine, were it not for the fact that the data are very exceptional, having a phenomenal mode.

The proper labeling of class intervals is important in connection with grouping. Class intervals of either grouped or ungrouped scores should be labeled by recording the lower and upper limits of the interval, e.g., 75.50–76.50, or by labeling the mid-point of the interval, e.g., 76.0. If the successive class intervals are the same the labeling of the mid-point is both clear and concise. A great deal of needless confusion is caused by improper labeling of intervals. The writer has found this especially true with reference to age data, such as the following:

Age	Height in Cm.	or Again,	Age	Score in Arithmetic Test
12	140		12	18.324
13	150		13	20.002
14	155		14	20.980
15	160		15	23.545

With data such as these it is a matter of sheer guess whether the scores correspond to mean ages of 12.0, 13.0, etc., or of 12.5, 13.5, etc. If a single score is recorded for a class interval it should universally be that of the mid-point of the interval, and in order to make it unambiguous the labeling figure should be carried one decimal further than the unit representing the class interval, e.g., if the above tables had read:

Age	Height in Cm.	Age	Score in Arithmetic Test
12.0	140	12.5	18.324
13.0	150	13.5	20.002
14.0	155	14.5	20.980
15.0	160	15.5	23.545

140 would have been taken as the mean height of individuals exactly twelve years old, etc., and no uncertainty would arise.

Section 12. The Median

The median of a series is the value of the mid-most measure, hence half the measures composing the series lie above it and half below.

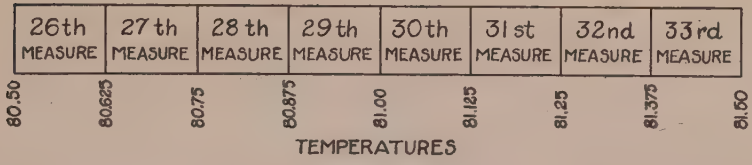
We will proceed to calculate the median of the daily maximum temperatures in New York City for July and August, 1917. The raw data are given in Table VIII. A hasty inspection shows that the lowest daily maximum temperature is 65° and the highest 98° and, a priori, knowing of no reason to expect that the distribution is skew it is assumed that the median lies about halfway between these two extremes. We will, therefore, make out a table of frequencies, as shown below:

f

Number of Days Having Temperatures Noted

Temperature	below	80	-|-|-|-|-	-|-|-|-|-	-|-|-|-|-	= 15	} 25
"	of	80	-|-|-|-|-	-|-|-|-|-		= 10	
"	"	81	-|-|-|-|-	| | |		= 8	
"	"	82	-|-|-|-|-			= 5	} 29
"	"	83	-|-|-|-|-	| |		= 7	
"	"	84	| |			= 2	
"	above	84	-|-|-|-|-	-|-|-|-|-	-|-|-|-|-	= 15	
						62	

Adding up measures from both ends, it is found that the median measure lies in the group with temperature 81°; or, since there are 62 measures, it lies halfway between the values of the 31st and 32d measures. As all measures from the 26th to the 33d inclusive are recorded as 81°, the 31st and 32d are so recorded and 81° may be taken as a rough approximation to the median. However, it is not to be presumed that the maximum temperatures on all of the eight days for which the temperature of 81° has been recorded were exactly 81.0°. It is more reasonable to consider that the average of these 8 temperatures was 81 and that they ranged all the way from 80.5 to 81.5. Furthermore, since this interval is small with reference to the entire range of temperatures, 34°, we may with satisfactory warrant consider that these 8 measures are evenly distributed over the interval 80.5–81.5, as shown in the diagram on page 55.



It is immediately seen that the temperature midway between the 31st and 32d measures is 81.25°. This is therefore the median sought.

This method is not the best possible, but gives a good determination for all practical purposes. For other methods see Bowley (1907). The best possible median is determined by mathematically fitting a curve to the observations and then integrating (or summing areas) from one end of the curve up to the point giving one half the total area. As thus determined the median is a function not only of position above or below a certain class value, but also of the distances of the measures above and below this median class, because the magnitude of each of the measures from the lowest to highest enters into the determination of the equation which fits the distribution.

Following in principle this integrating method, a median may be determined mechanically from a carefully plotted frequency polygon by the use of a planimeter. A guess is made as to the median and a perpendicular erected. The planimeter is run around the boundary of the area thus cut off and the result noted. If the area recorded by the instrument is not exactly one half the total area an adjusted guess as to the median is made and the process repeated. This may be continued until the desired degree of accuracy is obtained. Continuing the preceding illustration: If 63 days had been considered, and if the temperature of the added day had been greater than 81° there would have been one measure, the 32d, which would have had just as many measures below it as above, and the temperature corresponding to the middle of this mid measure, 81.3125°, would be the median. The median, or mid measure, may therefore be defined as the value of the $(N + 1)/2$ measure, but as the value of a measure is the value of its mid-point, this is equivalent to saying that the median is the limit of the range covered by $N/2$ measures

counted either down from the top or up from the bottom. The method pursued in the calculation of a median may be summarized and expressed in a formula as follows:

1. Arrange the measures in order of magnitude and list the frequencies for each class interval, grouping such intervals as are well below, or well above the median interval.

2. Let N = the total number of cases, i.e., the sum of the frequencies of all the classes.

3. Determine the class in which the $(N + 1)/2$ measure lies. If it lies between two classes, as sometimes happens when N is even, the common boundary of these two classes is the median and no further calculation is necessary. (The infrequent case when these two classes do not have a common boundary is treated in the next paragraph.)

4. Let f = the frequency of this class.

5. Let i = the class interval, or range covered by the median class.

6. Let F = the sum of the frequencies of all the classes below this class.

Let F' = the sum of the frequencies of all the classes above this class.

7. Let v = the value of the lower boundary of this class.

v' = the value of the upper boundary of this class.

8. Let Mdn = the median value. Then

$$\text{Mdn} = v + \frac{\frac{N}{2} - F}{f} i \qquad \text{(Median calculated from below up)} \ldots\ldots [2]$$

$$\text{Mdn} = v' - \frac{\frac{N}{2} - F'}{f} i \qquad \text{(Median calculated from above down)} [2\,a]$$

These two values of the median will be identical.

Using the first of these formulas to calculate the median of the maximum temperatures we have the following:

$N = 62$
$f = 8$
$i = 1$ (i.e., 1°)
$F = 25$ (frequencies below the median class)
$v = 80.5$ (lower boundary)

$$\text{Mdn} = 80.5 + \frac{31 - 25}{8} 1 = 81.25$$

Or again, using the second of these formulas and calculating from above down:

N, f, and i as above
$F' = 29$ (frequencies above the median class
$v' = 81.5$ (upper boundary)

$$\text{Mdn} = 81.5 - \frac{31 - 29}{8} 1 = 81.25$$

All cases have been covered by steps 1 to 8 except when the median lies between two classes which do not have a common boundary, as in the accompanying illustration: Here the $(N + 1)/2$ measure lies between classes c and e, but the upper limit of class c, 5.5, is not at the same time the lower limit of class e, 6.5. The median value might be considered to lie anywhere between 5.5 and 6.5, but the most reasonable procedure is to call it the average of these two values. The median is therefore $(5.5 + 6.5)/2 = 6.0$. With this understanding every distribution yields a single value for the median. If this value has been calculated from the bottom up it is well to check by calculation from the top down.

Frequencies	Scores	Classes
1	9	g
3	8	f
1	7	e
0	6	d
2	5	c
2	4	b
1	3	a
10		

Section 13. Percentiles

The median is the value below which 50 per cent of the measures lie. It is, therefore, the 50-percentile. Similarly the 10-percentile is the value below which 10 per cent of the measures lie, etc. The derivation which gave the formula for the calculation of the median may readily be generalized so as to provide a formula for the calculation of any percentile.

Let N = the total population.

Let P_p = the percentile, the value of which is to be calculated.

Let p = the proportion of cases having values smaller than P_p. Thus P_p is the 100 p-percentile. For example, if the 15-percentile is being considered, $p = .15$, and $P_{.15}$ is the symbol standing for the value of the 15-percentile.

Determine the class in which the 100 p-percentile, or the $(pN + \frac{1}{2})$ measure, lies.

Let f_p = the frequency in this class.

Let i_p = the interval or range covered by this class.

Let F_p = the sum of the frequencies in all the classes below this class.

Let v_p = the value of the lower boundary of this class interval.

Then:

$$P_p = v_p + \frac{pN - F_p}{f_p} i_p$$ (Value of a percentile — calculated from below up). . . . . [3]

This is the formula for the calculation of any percentile proceeding from small values of the variable to large values. If the calculation is from the other end of the distribution the formula is:

$$P_p = v'_p - \frac{(1 - p) N - F'_p}{f_p} i_p$$ (Value of a percentile — calculated from above down) . . . [3 a]

in which,

v'_p = the value of the upper boundary of this class interval

F'_p = the sum of the frequencies in all the classes above this class.

To insure accuracy it is well to calculate from below up and also from above down.

The same procedure as in the case of the calculation of the median is to be followed in the case of a percentile lying somewhere in a group with zero frequency.

For sake of illustration this formula will be used to calculate (a) the 50-percentile (the median), (b) the 25-percentile (the lower quartile), and (c) the 75-percentile (the upper quartile), for the temperature data.

(a) The Median (Mdn)

$$N = 62$$
$$p = .50$$
$$(.50)\ 62 + \tfrac{1}{2} = 31\tfrac{1}{2}$$

The $31\frac{1}{2}$ measure lies in the 81° class.

$$f_{.50} = 8$$
$$i_{.50} = 1$$
$$F_{.50} = 25$$
$$v_{.50} = 80.5$$
$$P_{.50} = 80.5 + \frac{(.50)\ 62 - 25}{8} 1. = 81.25$$

Note that in calculating from the top down $F'_{.50} = 29$, and $v'_{.50} = 81.5$.

(*b*) The lower quartile (L.Q.).

$$N = 62$$
$$p = .25$$
$$(.25)\,62 + \tfrac{1}{2} = 16.$$

The 16th measure lies in the 80° class.

$$f_{.25} = 10$$
$$i_{.25} = 1$$
$$F_{.25} = 15$$
$$v_{.25} = 79.5$$
$$P_{.25} = 79.5 + \frac{(.25)\,62 - 15}{10}\,1 = 79.55$$

Note that in calculating from the other end $F'_{.25} = 37$, and $v'_{.25} = 80.5$.

(*c*) The upper quartile (U.Q.).

$$N = 62$$
$$p = .75\ (.75)\,62 + \tfrac{1}{2} = 47.$$

The 47th measure lies in 84° class.

$$f_{.75} = 2$$
$$i_{.75} = 1$$
$$F_{.75} = 45$$
$$v_{.75} = 83.5$$
$$\therefore\ P_{.75} = 83.5 + \frac{(.75)\,62 - 45}{2}\,1 = 84.25$$

In calculating from above down $F'_{.75} = 15$, and $v'_{.75} = 84.5$.

The difference between the two quartiles is the interquartile range and of necessity 50 per cent of the cases lie in this range. In the problem in hand the interquartile range is 4.7° and indicates that one half of the days studied had maximum temperatures within 4.7° of each other.

The consideration of percentiles has been a diversion from the main purpose of this chapter, the study of averages, occasioned by their intimate connection with one of these averages, but we will here take up the main problem again in the study of the mode.

Section 14. The Mode

The mode is the value in a series at which the greatest frequency lies, or it is the place of densest frequency. In the case of Price Indexes, Table XV, this greatest frequency lay at "no change" in price, which is accordingly the mode.

In the case of College Marks, Chart X, a pronounced mode at 90 is shown by the raw data. However, such data have several modes and it is correct to speak of the distribution as multi-modal. If, from a priori consideration, it is thought that the minor modes are due to causes either chance or irrelevant with reference to the main trend, it is desirable to smooth them out and determine the one mode. In the case of College Marks the minor modes at 85, 80, 75, etc., are not due to chance but to psychological causes lying in the minds of instructors when called upon to grade individuals upon a finer scale than parallels their competency to make judgments. These modes at 85, 80, etc., would not be expected to vanish if the population were increased many fold, but the minor modes in the temperature data, Chart II at 83°, 85°, 75°, etc., are probably due to chance and would disappear if records for a number of years were taken, but the mode at 80° would probably remain, though it might shift slightly one way or the other. If one is studying temperatures this latter mode only is significant. If one is studying the distribution of talent of pupils, the major mode only of the College Marks distribution is wanted, while if one is studying the psychology of pedagogues the minor modes are very significant.

Assuming that the major mode only is sought we will consider its calculation. It is obvious that if the mode shown by the raw data is taken it will be very unreliable, for usually a change of but a measure or two will shift the mode, e.g., a shift of but a single measure in the temperature data from 80° to 81° would make it indeterminate whether the mode was 80° or 81° while a shift of two measures from 80° to 83° would shift the mode 3°. For this reason the mode is always determined from smoothed data if the raw data show irregularities in the vicinity of the mode.

The College Marks data have been smoothed by the moving

average method. (Sec. 6.) A perusal of Table XVIII shows that an unquestioned mode is not established by the class frequencies given by a moving average involving three classes. In that case modes exist at 86, 89, 91 and 94 — the largest of these being that at 89. When five class frequencies are averaged, modes appear at 88, 90 and 91 — the largest being at 88, so that the mode is still undetermined. When fifteen frequencies are averaged a single mode appears at 89, but the frequency of the 89 class is only .13 larger than that of the 90 class, out of a population of 773, so that the reliability of the determination is obviously not very great.

The distribution of frequencies given by averaging ten classes does establish the mode at 89.5 (the proof of this is left as an exercise) and accordingly 89.5 is the correct value to adopt as the mode.

The moving average method of determining the mode may be summarized as follows: Calculate smoothed class frequencies in the neighborhood of the mode, by means of a moving average involving a small number of intervals. Repeat the process, averaging greater and greater numbers of intervals, until a major mode with no minor modes in close proximity appears. The smallest grouping by which this major mode is obtained, gives the best result.

Another method for determining the mode follows from the relationship between the mean, median, and mode. Pearson has shown (1895) that in the case of his Type III curves the following relation holds:

Let Mo = mode, Mdn = median, M = mean, and σ = the standard deviation of the distribution (σ defined in the next chapter). Then

$$\text{Mo} = \text{M} - \frac{\text{M} - \text{Mdn}}{c}, \qquad \text{(The mode)} \ldots [4]$$

in which c is a magnitude differing slightly for different distributions and closely given by the equation

$$c = .3309 - \frac{.0846\,(\text{M} - \text{Mdn})^2}{\sigma^2 - 9\,(\text{M} - \text{Mdn})^2} \ldots\ldots\ldots\ldots [4a]$$

Therefore, knowing the mean, median and standard deviation, the mode may be calculated. Pearson's Type III curve is a skew curve limited at one end and unlimited at the other.

It is a very flexible curve and excellently represents a large number of skew distributions. If by inspection, a curve seems to approach a finite limit at one end, to be unlimited at the other, and if its kurtosis (see Sections 10 and 36) is not extreme, no serious error is likely to be introduced by assuming it to be a Type III curve.

Since the mean and median can be very reliably determined, the mode derived from them is a very much more stable measure than that as determined in the last section.

In case the distribution has a pronounced mode near the end at which it terminates, and a long and very thin tail at the other end, e.g., of the type of the distribution of incomes, it is well to use Formula [4 *a*], but for the great majority of skew distributions it is quite accurate enough to use $c = .33$. The mode is then given by equation:

$$\text{Mo} = \text{M} - 3.03\,(\text{M} - \text{Mdn}) \qquad \text{(The mode) [4 } b\text{]}.$$

Applying this method to the College Marks data for which 89.5 has already been found to be the mode, as calculated by means of a moving average, we have,

$$\begin{aligned} \text{M} &= 86.495 \quad \text{Calculated by formula [1]} \\ \text{Mdn} &= 87.690 \quad \text{“} \quad \text{“} \quad \text{“} \quad [2]. \\ \text{M} - \text{Mdn} &= 1.195 \\ \text{Mo} &= 86.495 - 3.03\,(-1.195) = 90.12 \end{aligned}$$

Of the two values obtained the greater credence should be given to 90.12. Using, instead of .33, the value of c as given by the full formula [4 *a*], leads to 90.13 as the mode; hence it is evident that the short formula is satisfactory for such a distribution as that of College Marks.

In handling distributions so decidedly skew that the skewness approaches 1.0, in which case $\sigma = 3(\text{M} - \text{Mdn})$, neither of the two formulas for calculating the mode from the mean and median can be used.

The three methods given, (*a*) graphic method of Section 7, (*b*) by smoothing the data, and (*c*) by derivation from the mean and median, are merely make-shifts if the student is able to avail himself of the precise determination resulting from mathematically fitting a curve to the data.

Section 15. The Harmonic Mean

Dunn's Wholesale Price Index is the cost of a year's supplies of a certain type. If the mean of the twelve of these indexes for a given year is calculated, it gives the mean cost of that year's supplies. But suppose instead of keeping the amount of goods constant and noting variability in price, the total cost had been kept constant and the variability in the amount of goods purchasable had been noted; how would one then proceed to obtain the mean cost of a given amount of goods? The following table, adapted from data given in Bradstreet's Journal, will serve to illustrate the problem:

Ruling Wholesale Prices, November 1

	1913	1914	1915	1916	1917	1918	
Pounds Sugar Bought for $1	23.0	18.5	19.4	13.33	11.9	11.11	(Designated as X measures)

Let it be desired to determine the mean price of a pound of sugar for the six years. We will first build up a table giving the cost per pound at the successive dates, by taking the reciprocals of the X measures as follows:

Ruling Wholesale Prices, November 1

	1913	1914	1915	1916	1917	1918	
Cost of Sugar in Dollars	.0435	.0540	.0515	.0750	.0840	.0900	(Designated as $\frac{1}{X}$ measures)

The mean of these measures is .06633 which accordingly is the mean cost of a pound of sugar for the six years. It is to be noted that if the mean of the X measures is found, 16.266, and the reciprocal taken, .06148, the same value is not obtained. The magnitude .06148 is not the mean price per pound — it is the reciprocal of the arithmetic mean number of pounds bought for $1, and a difficult measure to interpret, though not meaningless. The information of moment is the mean price per pound, or the reciprocal of this, the number of pounds which could be bought when paying the mean price per pound. This latter is the harmonic mean. In the case in hand it is the reciprocal

of .06633, or 15.08. Designating the harmonic mean by H.M. and employing the usual notation it is defined by the equation:

$$\text{H. M.} = \frac{1}{\frac{1}{N}\Sigma\frac{1}{X}} \qquad \text{(Harmonic mean)} \ldots \ [5]$$

In words: The harmonic mean is equal to the reciprocal of the mean of the reciprocals of the measures.

In deciding whether to use the arithmetic or the harmonic mean one should first decide which is properly the magnitude to remain constant (in the illustration, [*a*] the amount of sugar bought, or [*b*] the amount of money spent). There is seldom a doubt as to which should be the constant. If the data are recorded in such a manner that this appropriate item is constant, then the arithmetic mean is to be used. If the data, as recorded, make this item the variable, then the harmonic mean should be employed.

One further illustration may make this clearer. The following scores were made in a three-minute test in addition:

X: Numbers of Problems Completed	0	1	2	3	4	5	6	7	8	9	10	11	
f: Numbers of Pupils Making Scores Designated	0	0	1	0	4	7	10	8	3	2	2	0	Total = 37

The question should now be asked, Is the significant measure (*a*) the rate at which a pupil works a problem, or (*b*) the number of problems that he can work in a given time? The writer would judge that the rate at which the pupil works, or the number of minutes required to work one problem, is the more straightforward, readily comprehended and generally meaningful measure. Accepting this and noting that the data as recorded make the time element constant and not the number of problems worked, the harmonic mean is seen to be the proper mean to use.

If in this problem the arithmetic mean is calculated, there is a certain significance in it, but the reciprocal of this mean should not be compared with rate measures in which the number of problems is constant and the time allowed varies.

For discussion of the properties of an index number based upon the harmonic mean, see Fisher (1921).

Section 16. Geometric Mean

If the items in a series are so related (usually a temporal relationship) that the expression of each one in terms of the preceding one, i.e., relative to the preceding one, is the information required, then the averages thus far treated do not serve the purpose. These measures are, of course, ratios and the geometric mean is the significant average.

In Table XII, column two, are given the costs, on January 1 of successive years, of a year's supplies of certain common products. If the cost for each year is expressed in terms of the cost the preceding year, we have the following Table:

TABLE XXIV

Dunn's Wholesale Price Index for each Year Expressed as a Relative to the Preceding Year

1908	1.0561
1909	.9882
1910	1.1036
1911	.9325
1912	1.0724
1913	.9789
1914	1.0306
1915	.9971
1916	1.1087
	9)9.2681
	1.02979

If the mean advance per year is desired and the arithmetical mean, 1.02979, taken as the measure of it, serious error would be involved. The ratio of the basal year, 1907, with reference to itself is of course 1.00000, so that the mean advance as given by the arithmetic mean is .02979 and nine times this gives .2681, a measure for the advance over the entire period of nine years. That this is an incorrect measure is shown by the fact that the ratio of the prices in the last year to the basal year (137.666 ÷ 107.264) is 1.28343, showing that the actual advance is .28343. The reason for this discrepancy is that each advance is figured upon the preceding year as a base and not as a proportion of the price in the basal year. Strictly speaking 1907 is basal for 1908 only; 1908 being basal for 1909, etc. Accordingly 1.0561 × $107.264 gives the price for 1908. The price for 1908 times .9882 gives the price for 1909, or

.9882 × 1.0561 × $107.264; etc. Finally the products of all the nine ratios, 1.28343 times $107.264, gives $137.666, the price for 1916. In place of these nine different ratios whose product gives the ratio of the last year to the basal year, may be submitted a single mean ratio which, when multiplied by itself nine times, gives the same product. This is the geometric mean and, designating it by G.M. and the ratios for the separate years by $\rho_1, \rho_2, \rho_3, \ldots \rho_n$, it is defined by the equation:

$$\text{G. M.} = \sqrt[n]{\rho_1 \times \rho_2 \times \rho_3 \times \cdots \times \rho_n} \qquad \text{(Geometric mean)} \ldots [6]$$

It may be readily calculated by means of a Log Log slide rule or by means of logarithms as follows:

$$\log \text{G. M.} = \frac{\log \rho_1 + \log \rho_2 + \log \rho_3 + \cdots + \log \rho_n}{n} \text{ (Geometric mean) } [6\,a]$$

Using a slide rule the G. M. for the preceding data is found to be 1.0281. Using six place logarithms it is found to be 1.0282.

A check on these values is possible by taking the 9th root of the ratio of the 1916 price to the 1907 price. By logarithms this is found to be 1.02811. This figure means that on the average, wholesale prices increased 2.81 per cent each year, from 1907 to 1916.

The Index of Means, or of Sums

Another problem arises in connection with indexes which may be illustrated by the wage data in the last three columns of Table XII. The essential portions are copied below:

Chicago

	Union Wage per Hour		
	Painters	Linotype Operators	Carpenters
1907	50¢	50¢	56.3¢
1916	70	50	70
Same data expressed as ratios — 1907 as base			
1907	100	100	100
1916	140	100	124.334

Let us suppose that there are the same numbers employed from each of the unions, and let us designate this number by N. The question that concerns us is how to determine the average increase in wages. Does $\frac{140 + 100 + 124.334}{3} = 1.2144$, indicating an increase of 21.44 per cent, give it?

Bearing this in mind let us approach it by another method. The mean hourly wage in 1907 is $\left(\frac{N\,50 + N\,50 + N\,56.3}{3\,N}\right) =$ 52.10 cents, and in 1916 it is $\left(\frac{N\,70 + N\,50 + N\,70}{3\,N}\right) = 63.33$ cents. Dividing 63.33 by 52.10 the ratio of the mean wage in 1916 to that in 1907 is found to be 1.2156, giving an increase of 21.56 per cent. The two values found are not identical and it can be easily proven that in general they will not be, for, letting P, L and C equal the initial wages in the three unions respectively, and p, l, and c the ratios of the final wages to the initial wages in the three cases; then $p\,P$, $l\,L$, and $c\,C$ equal the final wages respectively; and, $\frac{NP + NL + NC}{3\,N} =$ the mean initial wage; also, $\frac{Np\,P + Nl\,L + Nc\,C}{3\,N} =$ the mean final wage; and the ratio of these two wages is $\frac{p\,P + l\,L + c\,C}{P + L + C}$. This is identical with $\frac{p + l + c}{3}$ only in case $P = L = C$, which in general is not the case. The fact that the initial wages were so nearly equal in the illustration accounts for the small difference in the two results.

We may therefore conclude that it is inaccurate to take the mean of ratios as equivalent to the ratio of the means (or sums) of final and initial scores.

Section 17. Weighting

If the numbers of workers in the three trades had been the same throughout and if because of considerations other than population the trades possessed importances W, w, w, then it would have been proper to multiply the wages by amounts equal or proportionate to W, w, w. This is "weighting."

The multiplying of a score by the number of cases having it has at times been called weighting, but in this text the term will be used to mean the multiplying of scores by amounts determined not at all, or not solely, by the population, but from other evidences of importance. (See Section 91.)

It is generally a difficult problem to determine just what constitutes proper weighting. When one is confronted with the problem of weighting measures which are to be combined and feels incompetent to accurately judge of their relative importances he is inclined to "solve" the problem by "not weighting at all." But the failure to assign weights is actually a very definite weighting — that of calling the units involved in the various measures of equal importance. This is not the same as saying that the failure to assign weights results in giving equal importance to the different items. This latter is not the case if the dispersions of the scores for the various items differ. This point, together with others involved in weighting, is treated at length in connection with partial correlation. It may certainly be said that, judging by the ordinary run of studies in economics and psychology, much more error has been committed by "not weighting at all" than by improper weighting.

PROBLEMS

1. Calculate the mode for the maximum temperature data of Table VIII. Is the short formula, in which $c = .33$, appropriate to use in this case?

2. Calculate the L. Q., Mdn. and U. Q. for the hypothetical distribution of incomes, comparing with graphic determinations (Problem 5, Chapter II).

Calculate the mean. Assume that the mean income for the highest income group is £21,000. Since these data have very irregular class intervals, in calculating the mean, great care must be taken in assigning ξ values to the different classes, no matter where the arbitrary origin is chosen. For this reason it will be more accurate and almost as short if the method given by Formula [1] is followed. The student may well make the calculation both ways to become familiar with the handling of irregularly grouped data.

Calculate the mode: (*a*) by finding the point of inflection in a smoothed ogive curve, (*b*) by deriving from the values of the mean and median, using $c = .33$ and (*c*) the same, using the full formula for c. In doing this take $\sigma = \frac{3}{4}$ the interquartile range.

3. The following three series are scores of individuals in three tests. They may be used as practice series for the calculation of *M*, Mdn., L. Q. and of constants treated of in subsequent chapters.

Practice Series

Individuals	Scores of a Class in Test 1	Scores of Same Class in Test 2	Scores of Same Class in Test 3
	Series 1	*Series 2*	*Series 3*
A	151	132	148
B	147	132	143
C	145	130	153
D	138	128	148
E	134	121	135
F	124	103	134
G	120	105	138
H	118	122	138
I	116	99	128
J	114	124	129
K	113	109	131
L	107	99	136
M	106	103	124
N	105	98	126
O	104	108	133
P	101	104	122
Q	100	115	137
R	99	111	119
S	98	107	121
T	96	92	124
U	89	96	118
V	87	94	126

4. Calculate the 5th, 10th, 15th, etc., percentiles for the scores in handwriting upon the Ayres and Thorndike scales, given in Table XXX, Section 34, and check answers against columns 1 and 2, Table XXXII, Section 35.

Group the Ayres data in 3's and the Thorndike data in 5's, calculate the same percentiles and check against answers in columns 3 and 4 of Table XXXII.

CHAPTER IV

MEASURES OF DISPERSION

Section 18. The Mean Deviation

Distributions having the same average may differ markedly in the spread of the measures composing them. The following two series of measures have the same mean, median and mode, but the scatter of the measures is very different:

7,	7,	7,	8,	8,	8,	8,	8,	9,	9,	9
1,	1,	1,	1,	3,	8,	13,	15,	15,	15,	15

The range in the first series is three, while in the second it is fifteen. If deviations from the mean, 8, are calculated, they run:

− 1,	− 1,	− 1,	0,	0,	0,	0,	0,	1,	1,	1	Sum = 6
− 7,	− 7,	− 7,	− 7,	− 5,	0,	5,	7,	7,	7,	7	Sum = 66

The means of these two series of deviations are of course zero if taken algebraically, but if taken absolutely, i.e., irrespective of sign, they are .545 and 6.0 respectively. These are the mean deviations.

The mean deviation may be defined as the sum of the absolute values of the deviations of the separate measures from the mean, divided by the population.

It can be calculated by the method of moments. Referring to Table XXII, columns four and five: If the deviations had been from the mean, 81.548 (in which case they would have been designated by x instead of by ξ) instead of from 80, a mere guess, the products $f \cdot x$, would have been slightly different from those recorded in column $f \cdot \xi$, and their sum, irrespective of sign, divided by their number, 62 would have been the mean deviation. Since, however, the calculation of deviations from the mean, 81.548, involves fractional or decimal magnitudes it is in practice inconvenient to determine the mean

deviation in this manner. Deviations from 81.548 run as given herewith in line (x):

(x): -16.548 -15.548 -14.548... -2.548 -1.548 $-.548$ $.452$ 1.452...16.452
(ξ): -15 -14 -13 ... -1 0 1 2 3 ...18

For purposes of comparison, the corresponding deviations from the arbitrary origin, 80, are given in line (ξ). It is seen that algebraically each ξ measure is 1.548 larger than the corresponding x measure. In absolute value all the ξ deviations up to and including those for class 80°, 25 in number, are 1.548 too small; those in class 81°, 8 in number, are .452 too large; and those in classes 82° and on, 29 in number, are 1.548 too large. Tabulated, the data show:

25 measures 1.548 too small
29 measures 1.548 too large

Excess of 4 measures 1.548 too large = excess positive moment of $4 \times 1.548 = 6.192$
Excess of 8 measures .452 too large = excess positive moment of $8 \times .452 = 3.616$
Total excess positive moment = 9.808

The sum of the moments as calculated from 80° is $89 + 185 = 274$, but this is too large by 9.808. Accordingly the sum of the deviations from the mean is 264.192 which, divided by 62, gives 4.26, the mean deviation sought.

The calculation, as shown, is cumbersome. A simple formula for the calculation of the mean deviation from the first moment about zero as an arbitrary origin is herewith derived.

Given the series 11, 12, 13, 13, 16. Mean = 13.0. The deviations of the successive measures from the mean are, -2, -1, 0, 0, 3 respectively, giving a mean deviation of 1.2. These deviations are (11–13), (12–13), (13–13), (13–13), (16–13), but since all are to be taken positively they must be written, (13–11), (13–12), (13–13), (13–13), (16–13). Using the usual notation we have:

$$\text{A. D.} = \frac{(M - X_1) + (M - X_2) + (X_3 - M) + (X_4 - M) + (X_5 - M)}{N}$$

$$= \frac{X_3 + X_4 + X_5 - X_1 - X_2 + M + M - M - M - M}{N}$$

If F = the number of measures lying below the mean (here 2), then it is seen that M enters in positively F times and negatively $(N-F)$ times and that the X's which are smaller than the mean enter in negatively (the sum of these may be represented by $\sum_{1}^{F} X$) and that those greater than the mean enter in positively (this sum may be represented by $\sum_{F+1}^{N} X$). Accordingly we have:

$$*\text{A. D.} = \frac{\sum_{F+1}^{N} X - \sum_{1}^{F} X + FM - (N - F)\,M}{N} \qquad [7]$$

Since, however,

$$\sum_{F+1}^{N} X - \sum_{1}^{F} X = \sum_{F+1}^{N} X + \sum_{1}^{F} X - 2\sum_{1}^{F} X = \sum_{1}^{N} X - 2\sum_{1}^{F} X$$

and since,

$$\sum_{1}^{N} X = NM$$

the formula becomes

$$\text{A. D.} = \frac{2}{N}\left(FM - \sum_{1}^{F} X\right) \quad \text{(Average deviation from the mean)} \quad [7\,a]$$

This is a very simple formula to use in connection with an adding machine. If the entries are not arranged according to magnitude add them on the machine and determine the mean, at the same time determining the population, N. Then add all the measures which are smaller than M, thus obtaining $\sum_{1}^{F} X$, at the same time determining the number of such measures, F. Thus two listings on an adding machine will yield the three important constants N, M and $A.D.$

If the measures are arranged according to magnitude a single listing will suffice, it only being necessary to take subtotals for each of the group frequencies in the neighborhood of the mean. For example the adding machine listing for the preceding series would be as shown herewith:

* This formula, with empirical proof, was independently discovered by two of the writer's students, Miss Elva Wald and Mr. John P. Herring.

	11
	12
2	23 *s*
	13
	13
4	49 *s*
	16
5	65 *t*

One would guess that the mean lay somewhere between 12 and 14 and would therefore take sub-totals after listing 12 and again after listing the 13's. Having $N = 5$ and the sum = 65, division gives the mean, 13.0. The listing shows that there are two measures below the mean and that their sum is 23, i.e., $F = 2$ and $\sum_{1}^{F} X = 23$. Thus immediately

$$\text{A. D.} = \tfrac{2}{5}\,(2 \times 13.0 - 23) = 1.2$$

The peculiar expedition of this formula should make it serviceable in large studies where time of computation is an important factor. It will shortly be shown that the probable error of the average deviation is but slightly greater than that of the standard deviation, so that unless the greatest accuracy is demanded, and unless the standard deviation is needed for such further purposes as use in correlation formulas, the average deviation will be found advantageous.

Returning to the Wald-Herring formula [7] it may be noted that if deviations around some point, P, other than the mean, be taken, and if F = the number of measures lying below this point, the formula becomes:

$$\text{A. D. around pt. } P = \frac{1}{N}\left[\sum_{F+1}^{N} X - \sum_{1}^{F} X + (2F - N)P\right]$$

(Average deviation around any point P) [8]

If $F = \dfrac{N}{2}$ then P is the median and the formula becomes:

$$\text{A. D. around Mdn} = \frac{1}{N}\left[\sum_{\frac{N+2}{2}}^{N} X - \sum_{1}^{\frac{N}{2}} X\right]$$

(Average deviation from the median) . . . . . [9]

Note that if N is odd, $\dfrac{N}{2}$ and $\left(\dfrac{N}{2} + 1\right)$ are fractional. In this

case it is necessary to add one half of the median measure in each summation. For the series 11, 12, 13, 13, 16;

$$\text{A. D. around Mdn} = \tfrac{1}{5}\,[(6.5 + 13 + 16) - (11 + 12 + 6.5)] = 1.20$$

This is the same as the average deviation from the mean for, in this particular problem, if measures are taken at their face value, the median and the mean coincide. Such measures as usually occur may, with insignificant error, regularly be taken at their face value in calculating the average deviation from the median, but they should not be so taken in calculating the median itself. The method already given in Section 12, based upon the assumption that the measures spread themselves evenly over the interval, is to be followed in calculating the median.

The mean deviation, unless stipulated to the contrary, is always calculated from the mean. It is at times desirable to calculate it from the median, in which case it should be definitely labeled "mean deviation from the median." A real reason for calculating it from the median exists in the fact that when so calculated it is smaller than when calculated from any other point, as can readily be shown:

Let ζ = a deviation from the median. Then the

$$\text{M. dev. from the Mdn} = \frac{\Sigma\,|\,\zeta\,|}{n}$$

Let ξ = a deviation from a point P which is Δ distance from the median; $\Delta <$ one class interval. Then $\xi = \zeta + \Delta$.

$$\text{M. dev. from } P = \frac{\Sigma\,|\,\xi\,|}{n} = \frac{\Sigma\,|\,\zeta\,| + F\Delta - (n - F)\,\Delta}{n}$$

Suppose Δ is positive, then P lies above the median and $F > (n - F)$ so that the above right hand member $= \dfrac{\Sigma\,|\,\zeta\,|}{n} +$ a positive magnitude. If Δ is negative, P lies below the median and $F < (n - F)$, so that the right hand member still $= \dfrac{\Sigma\,|\,\zeta\,|}{n} +$ a positive magnitude. Therefore, whether point P lies above or below the median the mean deviation from it is greater than $\dfrac{\Sigma\,|\,\zeta\,|}{n}$, the mean deviation from the median. The proving of

this same relation when $\Delta >$ one class interval can be readily accomplished and is left as an exercise. Accordingly the mean deviation is a minimum when taken from the median.

Section 19. The Quartile Deviation

A measure of dispersion may be obtained by taking the difference between any two percentiles. One such measure, the difference between the upper and lower quartiles, or the interquartile range, has already been mentioned. The most customary measure, however, is one half this measure, the semi-interquartile range, which for convenience and brevity is called the quartile deviation, and is designated by "Q." Using the usual notation for the upper and lower quartiles, we have:

$$Q = \frac{U.Q. - L.Q.}{2} \quad \text{(Quartile Deviation)} \ldots\ldots [10]$$

It is to be noted that the quartile deviation is not a deviation from any of the averages thus far considered. It is simply a measure indicative of dispersion. If thought of as a deviation at all it should be as one from a point midway between the upper and lower quartiles. A rather better way to interpret it is as one half the interquartile range, a range within which lie 50 per cent of the measures.

Section 20. The 10–90 Percentile Range

A range somewhat larger than the interquartile range has advantages over it and the quartile measure derived from it, as a measure of variability. I have shown (Kelley 1921 new) that for a normal distribution the interpercentile range having the minimal error is that between the 6.917 and the 93.083 percentiles. A range but slightly different from this and having nearly as great reliability is that between the 10th and 90th percentiles. This distance is called D and is given as the most serviceable measure of dispersion based upon percentiles.

$$D = P_{.90} - P_{.10} \quad \text{(10–90 percentile range)} \ldots\ldots [11]$$

Its calculation and interpretation are very simple, and as over 72 per cent more cases are required to secure as great reliability

in the quartile deviation, this measure of dispersion is recommended wherever percentiles are used. Its relationship, in case of a normal distribution, with other measures of dispersion is given in Section 31. For proof of the next ten formulas the reader is referred to the reference cited.

The standard error of D is given by formula [16] which in turn depends upon formulas [40], [43] and the following:

$$r_{P_p P_{p'}} = \sqrt{\frac{p}{q}\frac{q'}{p'}} \text{ in which } p < p' \quad \text{(The correlation between any two percentiles } P_p \text{ and } P_{p'}) \ldots\ldots [12]$$

$$\sigma_{P_p - P_{p'}} = \sqrt{\frac{Npq}{(y)^2} + \frac{Np'q'}{(y')^2} - \frac{2\,Npq'}{yy'}} \quad \text{(The standard error of an interpercentile range)} \ldots\ldots [13]$$

in which $p < p'$ and y is the ordinate of the curve at the percentile P_p, and similarly for y' and $P_{p'}$.

Assuming normality, formula [13] becomes

$$\sigma_{P_p - P_{p'}} = \frac{\sigma}{\sqrt{N}}\sqrt{\frac{pq}{(z)^2} + \frac{p'q'}{(z')^2} - \frac{2\,pq'}{zz'}} \quad \text{(Standard error of an interpercentile range in a normal distribution)} \ldots\ldots [14]$$

in which z and z' are ordinates as given in Table K–W for arguments of q and q'. If, further, percentiles equally distant from the ends of the distribution are calculated, $p = 1 - p'$ and formula [14] becomes

$$\sigma_{P_p - P_{(1-p)}} = \frac{\sigma}{z\sqrt{N}}\sqrt{2\,p\,(q-p)} \quad \text{(Standard error of a symmetrical interpercentile range in a normal distribution)} \ldots\ldots [15]$$

We now obtain for the standard error of the 10–90 interpercentile range

$$\sigma_D = \frac{\sigma}{\sqrt{N}}\,2.279224 \ldots\ldots [16]$$

Entering Table K–W with $q = .1$ we find that $x = 1.281552$. Thus $D = 2.563104\,\sigma$ which gives

$$\text{P. E.}_D = \frac{.599786 * D}{\sqrt{N}} \quad \text{(Probable error of } D) \ldots [16\,a]$$

This is a very convenient formula, as, for ordinary purposes, we may take

$$\text{P. E.}_D = \frac{.600\,D}{\sqrt{N}} \ldots\ldots [16\,a]$$

* On p. 744 of the reference cited (Kelley 1921 new) this value is incorrectly given as .6001

Two other constants which are of value in determining the type of a curve are Sk and Ku defined by the following equations:

$$\text{Sk} = P_{.50} - \tfrac{1}{2}(P_{.90} + P_{.10}) \quad \text{(A measure of skewness based on percentiles)} \quad [17]$$

The standard error of Sk is

$$\sigma_{\text{Sk}} = .59914\,\frac{D}{\sqrt{N}} \quad \text{(The standard error of the percentile measure of skewness)} \quad [18]$$

$$\text{Ku} = \frac{Q}{D} \quad \text{(A measure of kurtosis based on percentiles)} \quad [19]$$

The standard error is

$$\sigma_{\text{Ku}} = \frac{.27779}{\sqrt{N}} \quad \text{(The standard error or the percentile measure of kurtosis)} \quad [20]$$

For a symmetrical distribution Sk = o and for a mesokurtic distribution Ku = .26315. If a given distribution has a Ku > .26315 it is platykurtic and if < .26215 it is leptokurtic.

We thus see that the percentiles of a distribution may be used to answer some of the important questions of curve type. If populations are large, so that standard errors are small, resort to the longer though generally more accurate (not always, as it is dependent on curve type) methods of Chapter VII may frequently be avoided.

Section 21. The Standard Deviation

The standard deviation is far more universally significant than are any of the preceding. It is based upon the squares of the deviations from the mean, instead of upon the first powers as is the mean deviation. The exceptional advantages of this measure of dispersion will appear in connection with subsequent work. The standard deviation is defined as the square root of the mean of the squares of the deviations and is regularly designated by "σ." Unless otherwise stipulated deviations are always from the mean. Using the usual notation:

$$\sigma = \sqrt{\frac{\Sigma x^2}{n}} \quad \text{(The standard deviation of a distribution)} \quad [21]$$

This is a fundamental formula and should be recognized whether written as

$$\sigma^2 = \frac{\Sigma x^2}{n} \quad [21\,a]$$

or as,

$$\Sigma x^2 = n\sigma^2 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots [21\ b]$$

The calculation of the standard deviation for the temperature data of Table VIII is as follows:

TABLE XXV

Calculation of σ

Gross Score	Frequencies	Dev. from Arb. Orig.	First Moments	Second Moments	Second Moments from Mean
X	f	ξ	$f\cdot\xi$	$f\cdot\xi^2$	fx^2 $\quad A \quad B \quad C$
65	1	−15	−15	225	$1(-15-\delta)^2 = 1(15^2 - 2\delta[-15] + \delta^2)$
66	1	−14	−14	196	$1(-14-\delta)^2 = 1(14^2 - 2\delta[-14] + \delta^2)$
—					
69					
70	1	−10	−10	100	$1(-10-\delta)^2 = 1(10^2 - 2\delta[-10] + \delta^2)$
71	1	−9	−9	81	$1(-9-\delta)^2 = 1(9^2 - 2\delta[-9] + \delta^2)$
72					
—					
74	2	−6	−12	72	$2(-6-\delta)^2 = 2(6^2 - 2\delta[-6] + \delta^2)$
75	3	−5	−15	75	$3(-5-\delta)^2 = 3(5^2 - 2\delta[-5] + \delta^2)$
76	1	−4	−4	16	$1(-4-\delta)^2 = 1(4^2 - 2\delta[-4] + \delta^2)$
77	1	−3	−3	9	$1(-3-\delta)^2 = 1(3^2 - 2\delta[-3] + \delta^2)$
78	3	−2	−6	12	$3(-2-\delta)^2 = 3(2^2 - 2\delta[-2] + \delta^2)$
79	1	−1	−1	1	$1(-1-\delta)^2 = 1(1^2 - 2\delta[-1] + \delta^2)$
			−89		
80	10	0		0	$10(0-\delta)^2 = 10(0 - 2\delta[0] + \delta^2)$
81	8	1	8	8	$8(1-\delta)^2 = 8(1^2 - 2\delta[1] + \delta^2)$
82	5	2	10	20	$5(2-\delta)^2 = 5(2^2 - 2\delta[2] + \delta^2)$
83	7	3	21	63	$7(3-\delta)^2 = 7(3^2 - 2\delta)[3] + \delta^2)$
84	2	4	8	32	$2(4-\delta)^2 = 2(4^2 - 2\delta[4] + \delta^2)$
85	4	5	20	100	$4(5-\delta)^2 = 4(5^2 - 2\delta[5] + \delta^2)$
86	3	6	18	108	$3(6-\delta)^2 = 3(6^2 - 2\delta[6] + \delta^2)$
87	1	7	7	49	$1(7-\delta)^2 = 1(7^2 - 2\delta[7] + \delta^2)$
88	2	8	16	128	$2(8-\delta)^2 = 2(8^2 - 2\delta[8] + \delta^2)$
—					
90	1	10	10	100	$1(10-\delta)^2 = 1(10^2 - 2\delta[10] + \delta^2)$
—					
95	1	15	15	225	$1(15-\delta)^2 = 1(15^2 - 2\delta[15] + \delta^2)$
96	1	16	16	256	$1(16-\delta)^2 = 1(16^2 - 2\delta[16] + \delta^2)$
—					
98	2	18	36	648	$2(18-\delta)^2 = 2(18^2 - 2\delta[18] + \delta^2)$
99	—		—	—	
	62		185		
			96	2524	
			$\delta = 1.548$	40.710	$\Sigma\xi^2 \quad -2\delta\Sigma\xi \quad \Sigma\delta^2$

$$\sigma = \sqrt{40.710 - (1.548)^2} = 6.190$$

If the arbitrary origin, 80, had been the mean, the standard deviation would be given by $\sqrt{2524/62}$, but as the arbitrary origin is an amount δ, $(= \Sigma\xi/N = 96/62 = 1.548)$, below the mean, each ξ deviation is algebraically too large by the amount δ. Accordingly, if, in place of $\Sigma\xi^2$ we calculate $\Sigma(\xi - \delta)^2$ it will lead to the appropriate sum from which to calculate σ. Magnitudes $(\xi - \delta)^2$ are expanded and tabulated in the last three columns of Table XXV. It is immediately seen from the table, and is of course also apparent by squaring the binomial, that $\Sigma x^2 = \Sigma(\xi - \delta)^2 = \Sigma\xi^2 - \Sigma\, 2\,\delta\xi + \Sigma\delta^2$. Since δ is a constant and does not vary from class to class $\Sigma\, 2\,\delta\xi = 2\,\delta\Sigma\xi$ and similarly $\Sigma\delta^2 = N\delta^2$ (here $= 62 \times 1.548^2$). The summation $\Sigma\xi$ has already been obtained in summing the first moments and, from the definition of δ, $\Sigma\xi = N\delta$. Accordingly $\Sigma x^2 = \Sigma\xi^2 - 2\,N\delta^2 + N\delta^2$, and

$$\sigma = \sqrt{\frac{\Sigma\,\xi^2}{N} - \delta^2} \qquad \text{(The standard deviation of a distribution calculated from an arbitrary origin)} \quad [22]$$

The symbol δ, usually standing for a small magnitude, should not be so interpreted here, for the formula is rigorously exact whether the arbitrary origin differs from the mean by a fraction of a unit or a large number of units.

The square of the standard deviation, σ^2, is frequently an essential constant. It is designated by u_2, meaning the second moment about the mean. Without further explanation the meanings of the various moments, all taken from the mean, will be understood from the following equations, in which x, as usual, stands for a deviation from the mean:

$$\text{The first moment, } \mu_1 = \frac{\Sigma\, x}{N} = 0 \qquad [23]$$

$$\text{The second " } \mu^2 = \sigma^2 = \frac{\Sigma\, x^2}{N} \qquad [23\,a]$$

(Definition of the moments)

$$\text{The third " } \mu_3 = \frac{\Sigma\, x^3}{N} \qquad [23\,b]$$

$$\text{The fourth " } \mu_4 = \frac{\Sigma\, x^4}{N} \qquad [23\,c]$$

etc

If deviations from an origin, P, δ distance from the mean, O,

are calculated, then $O - P = \delta$, and $x = \xi - \delta$, and the following relationships hold:

$$\mu_1 = \frac{\Sigma (\xi - \delta)}{N} = \frac{\Sigma \xi}{N} - \delta$$

$$\mu_2 = \frac{\Sigma (\xi - \delta)^2}{N} = \frac{\Sigma \xi^2}{N} - \delta^2$$

$$\mu_3 = \frac{\Sigma (\xi - \delta)^3}{N} = \frac{\Sigma \xi^3}{N} - 3\delta \frac{\Sigma \xi^2}{N} + 3\delta^2 \frac{\Sigma \xi}{N} - \delta^3 = \frac{\Sigma \xi^3}{N} - 3\delta \frac{\Sigma \xi^2}{N} + 2\delta^3$$

$$\mu_4 = \frac{\Sigma (\xi - \delta)^4}{N} = \frac{\Sigma \xi^4}{N} - 4\delta \frac{\Sigma \xi^3}{N} + 6\delta^2 \frac{\Sigma \xi^2}{N} - 4\delta^3 \frac{\Sigma \xi}{N} + \delta^4$$

$$= \frac{\Sigma \xi^4}{N} - 4\delta \frac{\Sigma \xi^3}{N} + 6\delta^2 \frac{\Sigma \xi^2}{N} - 3\delta^4$$

$\mu_5 =$ etc.

If $\bar{\mu}_1$, $\bar{\mu}_2$, etc., stand for the moments around the arbitrary origin the above equations may be more simply written:

$$\left.\begin{aligned} \delta &= \frac{\Sigma \xi}{N} = \bar{\mu}_1 \\ \mu_1 &= \bar{\mu}_1 - \bar{\mu}_1 \\ \mu_2 &= \bar{\mu}_2 - \bar{\mu}^2_1 \\ \mu_3 &= \bar{\mu}_3 - 3\bar{\mu}_2\bar{\mu}_1 + 2\bar{\mu}^3_1 \\ \mu_4 &= \bar{\mu}_4 - 4\bar{\mu}_3\bar{\mu}_1 + 6\bar{\mu}_2\bar{\mu}^2_1 - 3\bar{\mu}^4_1 \end{aligned}\right\}$$ (The moments about the mean determined from those about any arbitrary origin) $\left\{\begin{aligned} &[24] \\ &[24\,a] \\ &[24\,b] \\ &[24\,c] \\ &[24\,d] \end{aligned}\right.$

etc.

The following formulas give the same results and are usually the more serviceable,

$$\left.\begin{aligned} \mu_1 &= 0 \\ \mu_2 &= \bar{\mu}_2 - \bar{\mu}^2_1 \\ \mu_3 &= \bar{\mu}_3 - 3\mu_2\bar{\mu}_1 - \bar{\mu}^3_1 \\ \mu_4 &= \bar{\mu}_4 - 4\mu_3\bar{\mu}_1 - 6\mu^2\bar{\mu}^2_1 - \bar{\mu}^4_1 \end{aligned}\right\}$$ (Moments about the mean determined from the moments about any arbitrary origin) $\left\{\begin{aligned} &[25] \\ &[25\,a] \\ &[25\,b] \\ &[25\,c] \end{aligned}\right.$

etc.

It is sometimes desirable to determine the moments from some arbitrary origin knowing them from the mean. Solution of the preceding formulas gives:

$$\bar{\mu}_n = \mu_n + n\,\mu_{n-1}\bar{\mu}_1 + \frac{n(n-1)}{2!}\mu_{n-2}\bar{\mu}^2_1 + \frac{n(n-1)(n-2)}{3!}\mu_{n-3}\bar{\mu}_1{}^3 + \cdots$$

(Moments about an arbitrary origin determined from moments about the mean) . . . [26]

In case the grouping is not fine a small correction to the μ's as given in formulas [68] is necessary.

We may now investigate some of the properties of the standard deviation. Let us compare the magnitudes of two standard deviations; (*a*) taken from the mean, O, and (*b*) from a point, P, δ distance from the mean. $O - P = \delta$, and $x = \xi - \delta$. Let σ = the standard deviation from O and s = the standard deviation from P:

$$\sigma^2 = \frac{\Sigma x^2}{N}$$

$$s^2 = \frac{\Sigma \xi^2}{N} = \frac{\Sigma (x + \delta)^2}{N} = \frac{\Sigma x^2}{N} + 2\delta \frac{\Sigma x}{N} + \frac{\Sigma \delta^2}{N}, \text{ and since } \frac{\Sigma x}{N} = 0$$

Hence

$$s^2 = \sigma^2 + \delta^2 \quad \ldots\ldots [27]$$

or

$$s = \sqrt{s^2 + \delta^2}$$ (Standard deviation about an arbitrary origin determined from the standard deviation about the mean) $\ldots\ldots [27\,a]$

Since δ, whether positive or negative, enters into this expression as a square, $s^2 > \sigma^2$; in other words, the standard deviation is a minimum when taken from the mean. This is a very important property of the mean.

Formula [24] for μ_2 gives the standard deviation squared in terms of the moments about an arbitrary origin. Formula [27] for s^2 gives the standard deviation squared from an arbitrary origin in terms of the second moment around the mean and the distance between the mean and arbitrary origin. It should, however, be noted that neither of these formulas gives the standard deviation around a second arbitrary origin in terms of the moments around a first arbitrary origin. This problem may readily be solved; if P and Q are the second and first origins and if ξ and ζ are deviations and s and S standard deviations around these origins respectively, we have:

$$P - Q = \Delta$$

$$\xi = \zeta - \Delta$$

$$s^2 = \frac{\Sigma \xi^2}{N} = \frac{\Sigma (\zeta - \Delta)^2}{N} = \frac{\Sigma \zeta^2 - 2\Delta\Sigma\zeta + N\Delta^2}{N} = S^2 + \Delta^2 - 2\Delta\overline{\mu}_1$$

(Relation between standard deviations about two arbitrary origins) $\ldots\ldots [28]$

Expressed in words: if moments around any two origins are taken, the second moment around the second origin equals the second moment around the first origin plus the square of

the difference between the origins minus twice the product of the difference (taking the second origin minus the first) and the first moment around the first origin.

The formula as written is to be used in determining the second moment around the "second" origin when the moments around the "first" origin are known.

Section 22. The Standard Error of the Mean

If it is desired to determine the reliability of the mean it is necessary to have an estimate of how a number of equally excellent, i.e., similarly derived, means distribute themselves — that is, a new distribution is to be conceived with the means themselves as the gross scores. The standard deviation of these means is indicative of the precision of any one of them. If this distribution of means has a very small spread, or standard deviation, then any one of them is a good measure, good in the sense that it is a close approximation to the mean of all the means. We thus need σ_M, the standard deviation of the means. If there are M sets of N measures each, and if the mean of the MN (where MN equals a very large number) measures, i.e., the mean of the means, is the true value, or true origin, then x stands for a deviation of a measure from this origin and $\frac{x_1 + x_2 + \ldots . x_N}{N}$, the mean of one set of N measures, is expressed as a deviation from this same origin. The standard deviation of such means is σ_m, the standard deviation sought. The standard deviation of the distribution of measures from the mean of the N measures will not be identical with the standard deviation of the same measures from the origin as here defined, but the difference may be expected to be negligibly small if N is larger than 25, which we shall assume to be the case in this derivation. We will designate the standard deviation of the original measures by σ. We have:

$$\sigma^2_M = \frac{\Sigma\left(\frac{x_1 + x_2 + \cdots x_N}{N}\right)^2}{M}, \text{ or}$$

$$MN\sigma^2_M = \Sigma\left(\frac{x^2_1 + x^2_2 + \cdots x^2_N + 2\,x_1x_2 + 2\,x_1x_3 + \cdots 2\,x_1x_N + 2\,x_2x_3 + \cdots + 2\,x_{N-1}x_N}{N}\right)$$

However, $\left(\frac{x^2_1 + x^2_2 + \cdots x^2_N}{N}\right) = \sigma_2$, and as Σ designates a summation of M such magnitudes, $\Sigma\left(\frac{x^2_1 + x^2_2 \cdots x^2_N}{N}\right) = M\sigma^2$.

Also $2\,x_1x_2 + 2\,x_1x_3 + \cdots$ may be rewritten, $x_1x_2 + x_1x_3 + \cdots x_1x_N + x_2x_1 + x_2x_3 + \cdots x_2x_N + x_3x_1 + x_3x_2 + x_3x_4 + \cdots x_3x_N + \cdots$, which, if S_1, S_2, $\cdots$ stand for summations of $N - 1$ terms each, is $= x_1S_1x + x_2S_2x + \cdots x_NS_Nx$. Each of these S summations is closely equal to zero. [Product theorem, see Section 23.] Since these summations are at times small positive and at other times small negative magnitudes and since $x_1\,x_2 \cdots$ are likewise both positive and negative and are entirely independent of the S's, it is clear that the whole expression, $(x_1S_1 + x_2S_2 + \cdots x_NS_N)$ does not vary from zero by but a small amount and is negligible in comparison with the sum of the square terms. The equation may then be written:

$$\text{V}M\sigma^2{}_M = M\sigma^2, \text{ or}$$

$$\sigma_M = \frac{\sigma}{\sqrt{n}} \quad \text{(Standard error of the mean) [29]}$$

This is a fundamental relation applicable when $n > 25 \cdots$ Expressed in words: The standard deviation of the mean equals that of the gross scores divided by the square root of the population.

Any measure whatsoever may be thought of as one of a distribution, the variability of the distribution being an indication of the error involved when any single measure of the distribution, taken at random, is chosen as the value of the thing measured. Thus when a measure is taken as the best obtainable value the standard deviation of just such measures as the one taken is the standard error. Thus the "standard error" of a measure and the "standard deviation" of such measures are synonymous expressions. The relation between the standard error and the probable error as derived in Section 28 is

Probable error = .6744898 standard error [Formula 33 of Sec. 27].

Section 23. The Standard Error of Any Moment

The product theorem used in the preceding derivation may be stated:

> The sum of products of measures which are independent of each other and whose means are zero, equals zero. [Product theorem]

This theorem, only roughly proven above, will later, in connection with the subject of correlation, be seen to be a necessary consequence of independence between measures. By utilizing it we may determine the standard deviation of any moment, μ_n, in a manner very similar to that in which we have determined the standard deviation of the first moment, μ_1, the mean.

Consider a population composed of M sets of N measures each. The n'th moment of the total population is, if Σ indicates a summation of M terms and S a summation of N terms:

$$\mu_n = \frac{\Sigma\,(Sx^n)}{MN}$$

The deviation from this value of a determination based upon one set of N measures is:

$$\left[\frac{Sx^n}{N} - \frac{\Sigma\,(Sx^n)}{MN}\right] = \left[\frac{Sx^n}{N} - \mu_n\right] = \left[\frac{S\,(x^n - \mu_n)}{N}\right]$$

This is a small magnitude. The sum of M such would of course be zero, but the sum of the squares would not, as there would then be no negative terms. Accordingly the standard deviation desired is:

$$\sigma_{\mu n} = \frac{\sqrt{\Sigma\left[\frac{S\,(x^n - \mu_n)}{N}\right]^2}}{M}$$

$S(x^n - \mu_n) = (x^n{}_1 - \mu_n) + \cdots (x^n{}_N - \mu_n) = \delta_1 + \delta_2 + \cdots \delta_N$, let us say. Then $MN\ \sigma^2{}_{\mu_n} = \frac{1}{N}\ \Sigma\ [S\delta]^2$ in which $[S\delta]^2 = S\delta^2 + 2\ S'\ \delta_p\delta_q$, where S' = a summation of $\frac{N(N-1)}{2}$ terms which approaches zero according to the theorem just stated. Accordingly,

$$MN\sigma^2{}_{\mu_n} = \frac{1}{N}\,\Sigma\,S\delta^2 = \frac{1}{N}\,\Sigma'\delta^2,$$

in which Σ' indicates a summation of MN terms.

Replacing the δ's by the equivalent binomials, we have:

$$MN\sigma^2\mu_n = \frac{1}{N}\Sigma'(x^{2n} - 2\mu_n x^n + \mu^2 n), \text{ which, since } 2\mu_n \Sigma' x^n = 2MN\mu^2 n$$

$$= \frac{1}{N}(MN\mu_{2n} - MN\mu^2 n)$$

$$\sigma_{\mu_n} = \sqrt{\frac{\mu_{2n} - \mu^2 n}{N}} \quad \text{(Standard error of any moment)} \ldots\ldots\ldots [30]$$

It is thus seen that the standard error of any moment is determined when that moment, the moment twice as large, and the population are known. It is to be noted that this formula is entirely general and does not depend upon having a symmetrical distribution. It only requires that the populations dealt with shall not be small.

Applying this formula to the determination of the standard deviation of the mean, $n = 1$, and we have:

$$\sigma m = \sigma\mu_1 = \sqrt{\frac{\mu_2 - \mu^2 1}{N}} \quad \text{(Standard error of the mean) [29 } a\text{]}$$

This is the general formula. It may be written more simply for it has already been pointed out that $\mu_1 = 0$, and $\mu_2 = \sigma^2$, so that the equation becomes:

$$\sigma m = \frac{\sigma}{\sqrt{N}} \quad \text{(Standard error of the mean)} \ldots [29]$$

This, of course, is identical with that previously derived.

We may determine the standard error of the standard deviation, but shall first need that of the standard deviation squared, μ_2: By formula [30] we have

$$\sigma_{\mu_2} = \sqrt{\frac{\mu_4 - \mu^2 2}{N}} \quad \text{(Standard error of the second moment)} \ldots [31]$$

It remains to determine what is the square root of a quantity corresponding to a given deviation in the quantity itself. Consider the magnitudes μ_2 and $(\mu_2 + \Delta)$ and also $\sqrt{\mu_2}$ and $\sqrt{\mu_2 + \Delta}$ or their equals σ^2 and $(\sigma^2 + \Delta)$ and also σ and $\left(\sigma + \frac{\Delta}{2\sigma} - \frac{\Delta^2}{8\sigma^3} + \cdots\right)$. (This latter after expansion of the radical by the binomial theorem.)

It is seen that corresponding to a small error Δ in σ^2, there is an error

$$\left(\frac{\Delta}{2\sigma} - \frac{\Delta^2}{8\sigma^3} + \cdots\right)$$

in σ. However, in all ordinary situations, $\Delta^2/8\sigma^3$ and higher terms are negligible in comparison with $\Delta/2\sigma$, so that we have:

$$\sigma_\sigma = \frac{1}{2\sigma}\sqrt{\frac{\mu_4 - \mu_2^2}{N}}$$

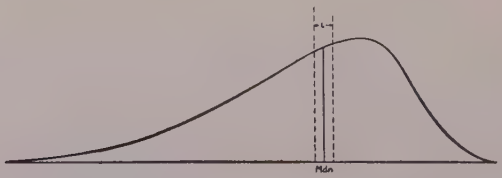
$$= \frac{1}{2\sigma}\sqrt{\frac{\mu_4 - \sigma^4}{N}} \quad \text{(Standard error of the standard deviation)} \ldots [32]$$

Utilizing formula [51] of Section 26 we have

$$\sigma_\sigma = \frac{\sigma}{\sqrt{2N}} \quad \text{(Standard error of the standard deviation in a normal distribution)} \ldots [32\,a]$$

Section 24. The Standard Error of a Class Frequency; of the Median; and of a Percentile

The deviation in the value of the median is a function of the deviation in the frequencies below, or above it. Consider the accompanying graph to represent the distribution of certain scores in the case of a very large population. If Δ frequencies are transferred from below Mdn, the median point, to above it, the median would be shifted up. The amount of this shifting may be readily determined.



Let f = the frequency in a small interval of range, i, near the center of which is the median.

Then the new median has been shifted an amount $i(\Delta/f)$ above the old median, assuming that the frequencies in the interval i distribute themselves in a rectangular manner. The fact that this assumption is not the most reasonable which can ordinarily be made has entirely insignificant influence in case distributions do not show very exceptional rates of change in

the vicinity of the median and in case populations are not small, let us say not less than 25.

It is thus seen that corresponding to a change Δ in the number of frequencies below the median, there is a definitely established change in the median. The standard error of the median may therefore be written,

$$\sigma_{\text{Mdn}} = \sigma_{\Delta} \frac{i}{f} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots [33]$$

It only remains to calculate the standard deviation of the Δ's and substitute in the above expression in place of σ_{Δ} to have the standard error of the median.

In drawing a sample of n measures from the total population, in which the chance of each measure lying below the median is one half, we will call those which lie below the median successes and those above failures and we will let F equal the number of successes. If two scores are drawn ($n = 2$) then the chance of both being successes; of the first being a success and the second a failure; of the first a failure and the second a success or of both being failures is $[(1/2)\,(1/2)]$ in each instance. Each of these is equally likely to occur, so that if a large number, N, of such samplings of two are made we have the following distribution of successes, or of frequencies lying below the median:

Successes	Frequencies
0	$N\frac{1}{2} \times \frac{1}{2} = N\frac{1}{4}$
1	$N2 \times \frac{1}{2} \times \frac{1}{2} = N\frac{2}{4}$
2	$N\frac{1}{2} \times \frac{1}{2} = N\frac{1}{4}$

That is, one fourth of the samplings will show no measures in this category (below the median), one half will show one measure in it, and one fourth will show two measures in it.

If three scores are drawn at a time there is just one permutation yielding three successes, three permutations yielding two successes and one failure, three yielding one success and two failures, and one yielding three failures, so that we have the following distribution:

Successes	Frequencies
0	$N\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = N\frac{1}{8}$
1	$N3 \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = N\frac{3}{8}$
2	$N3 \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = N\frac{3}{8}$
3	$N\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = N\frac{1}{8}$

That is, 1/8 of the samplings will show zero successes, 3/8 one success, 3/8 two successes, and 1/8 three successes.

If four are drawn ($n = 4$) the frequencies will run $N(1/16)$, $N(4/16)$, $N(6/16)$, $N(4/16)$, $N(1/16)$, and in general, if n are drawn at a time the frequencies will be given by the coefficients of the successive terms of the binomial $N(.5 + .5)^n$. Dropping N, which is a constant throughout, the general distribution may then be written:

Successes in Drawings of n at a Time	Frequencies
0	$1\ (\frac{1}{2})\ n$
1	$n\ (\frac{1}{2})\ n$
2	$\frac{n\ (n-1)}{2}\ (\frac{1}{2})\ n$
3	$\frac{n\ (n-1)\ (n-2)}{1 \times 2 \times 3}\ (\frac{1}{2})\ n$
etc.	etc.

Starting with this distribution we could readily determine its mean and standard deviation, but as it is just a special case of the more general problem in which the chance of success for any single drawing is p (p not necessarily $\frac{1}{2}$) this latter will be attacked.

Let p = the chance of success and q that of failure. Then

$$p + q = 1 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots [34]$$

Following the same argument as for $p = q = .5$, the distribution of successes when n at a time are drawn becomes:

Successes in n Drawings	Frequencies
0	$1 \quad q^n$
1	$n\, q^{n-1}\, p$
2	$\frac{n\ (n-1)}{2}\ q^{n-2}\, p^2$
3	$\frac{n\ (n-1)\ (n-2)}{1 \times 2 \times 3}\ q^{n-3}\, p^3$
etc.	etc.

We will now proceed to calculate the standard deviation of these numbers of successes by calculating the second moment from the point "zero successes," and then transferring to the mean by the aid of formula [22].

Successes in Drawings of n at a Time X	Frequencies f	fX
0	q^n	0
1	npq^{n-1}	npq^{n-1}
2	$\frac{n(n-1)(n-2)}{1 \times 2} p^2q^{n-2}$	$np(n-1)pq^{n-2}$
3	$\frac{n(n-1)(n-2)}{1 \times 2 \times 3} p^3q^{n-3}$	$np\frac{(n-1)(n-2)}{1 \times 2} p^2q^{n-3}$
4	$\frac{n(n-1)(n-2)(n-3)}{1 \times 2 \times 3 \times 4} p^4q^{n-4}$	$\frac{np(n-1)(n-2)(n-3)}{1 \times 2 \times 3} p^3q^{n-4}$
etc.	etc.	etc.

$$\Sigma f = (p+q)^n = 1 \qquad \Sigma fX = np(p+q)^{n-1} = np$$

Therefore $\bar{\mu}_1 = \frac{np}{1} = np$ . . . . [35]

fX^2

0

$$np(n-1)pq^{n-2} + np(n-1)pq^{n-2}$$

$$np\frac{(n-1)(n-2)}{1 \times 2} p^2q^{n-3} + np(n-1)(n-2)p^2q^{n-3}$$

$$np\frac{(n-1)(n-2)(n-3)}{1 \times 2 \times 3} p^3q^{n-4} + np\frac{(n-1)(n-2)(n-3)}{1 \times 2} p^3q^{n-4}$$

etc.

$$\Sigma fX^2 = np(p+q)^{n-1} + np^2(n-1)(p+q)^{n-2} = np + n^2p^2 - np^2 = \bar{\mu}_2$$

Therefore $\mu_2 = npq$, and $\sigma = \sqrt{npq}$ . . . . . [36]

The third and fourth moments, derived by the same process, are:

$$\mu_3 = npq(q-p) \text{ . . . . . [37]}$$

$$\mu_4 = npq[1 + 3(n-2)pq] \text{ . . . . . [38]}$$

They are recorded here for future reference, but are not used in the immediate problem, — the calculation of the standard error of the median.

The magnitude μ_2 is the standard deviation squared of the sum of the frequencies in a category for which the chance of each of the separate measures being in the category is p. Thus if N (instead of n as above) equals the size of the sample drawn, F the frequency in a certain category, p the likelihood of the measure lying without it, then

$\sigma_F = \sqrt{Npq}$ (The standard deviation of the frequency in a given category) . . . [39]

If the proportion in a category instead of the gross frequency is considered we have

$$p = \frac{F}{N} \text{ and } \sigma_p = \frac{1}{N}\sigma_F, \text{ so that finally}$$

$$\sigma_p = \sqrt{\frac{pq}{N}} \quad \text{(The standard deviation of a proportion)} \ldots [40]$$

This is the basic formula underlying the theory of contingency, i.e., the statistics of categories.

We may use this general result in determining the standard deviation of the frequencies below the median. In this case $p = q = \frac{1}{2}$, so that

$$\sigma_F = \frac{\sqrt{N}}{2}$$

This is the standard deviation of the Δ's, required to determine the standard error of the median. Substituting in [33]

$$\sigma_{\text{Mdn}} = \frac{i\sqrt{N}}{2f} \quad \text{(The standard error of the median)} \ldots [41]$$

By parity of reasoning the standard error of any percentile may be found. Using the same notation as in Section 13, it is

$$\sigma_{P_p} = \frac{i_p\sqrt{Npq}}{f_p} \quad \text{(The standard error of a percentile)} \ldots\ldots [42]$$

Formula [42] is ordinarily the one needed, but for certain problems the existence or assumption of normality permits the use of the following (Kelley, 1921, new);

$$\sigma_{P_p} = \frac{\sigma}{z}\sqrt{\frac{pq}{N}} \quad \text{(The standard error of a percentile of a normal distribution)} \ldots\ldots\ldots\ldots\ldots [43]$$

in which σ is the standard deviation of the distribution and z the ordinate corresponding to q as given in Table K–W.

A precaution is necessary in using formulas [41] and [42] in that, theoretically, f is the frequency in the interval i in the case of a very large population. A single class frequency for ordinary finite populations is a quite unstable magnitude, so that in determining the class frequency for f it is well to smooth the curve in the neighborhood of the percentile by averaging the three or five class frequencies nearest to it. The exact number to be averaged depends upon local periodicity and the

total population, but as a general rule for populations less than 200 it is advisable to average such a number as extend over approximately 1/8 of the total range. For larger populations a smaller number of intervals may be averaged. It is obvious that the same result is accomplished if the frequencies in a small number of neighboring intervals are added to give the f, and the total range covered by these intervals taken as the i, used in the formulas.

The standard errors of the two most important averages have been determined. That for the mode, except when calculated by determining the equation of the curve which fits the data, is known to be very high. No simple formula for its determination is available.

In order to compare the reliabilities of different averages we will calculate the standard errors of the mean and of the median for the temperature data of Table VII.

$$M = 81.55;\ \text{Mdn} = 81.25;\ \sigma = 6.19,\ N = 62$$

$$\sigma_M = \frac{6.19}{\sqrt{62}} = .786.$$

To compare with this, the standard error of the median will be calculated, using five different intervals in the neighborhood of the median.

(*a*) $i = 1$. f of interval, 80.5°–81.5°, $= 8$. $\sigma_{\text{Mdn}} = \frac{\sqrt{62}}{2} \times \frac{1}{8} = .493$

(*b*) $i = 2$. f of interval, 80.5°–82.5°, $= 13$. $\sigma_{\text{Mdn}} = \frac{\sqrt{62}}{2} \times \frac{2}{13} = .606$

(*c*) $i = 3$. f of interval, 79.5°–82.5°, $= 23$. $\sigma_{\text{Mdn}} = .514$

(*d*) $i = 4$. f of interval, 79.5°–83.5°, $= 30$. $\sigma_{\text{Mdn}} = .525$

(*e*) $i = 5$. f of interval, 78.5°–80.5°, $= 31$. $\sigma_{\text{Mdn}} = .636$

It is well-nigh impossible to say which of these five values is the most reliable, but since the population is only 62, the last value, .636, based upon an interval which is 1/7 of the range is rather to be preferred to any of the others. Accepting it as the best value it is seen that the median has a smaller standard error than the mean. This means that, if this sample of 62 is truly representative of the distribution of temperatures, the median of the distribution can be determined with greater

accuracy than can the mean, and that accordingly the median is preferable in this instance to the mean, as a measure of central tendency. Other considerations may enter in, such as, for example, the desirability of combining different sets of data, calculating correlations, etc., in which case the mean should always be used, as it permits of such statistical treatment whereas the median does not; but if such considerations are not present the *proper average to use is the one which is the most reliable*. It is thus seen that the all too customary choice of an average "because of the nature of the distribution" should give way to a choice based upon rigorous statistical considerations as to reliability. Having decided upon an average the appropriate measure of dispersion follows as a consequence — the quartile deviations or preferably D, the 10–90 percentile range, should be used with the median, and the standard deviation with the mean. The standard deviation is much the more reliable of these two measures of dispersion for all ordinary uni-modal distributions, even though they be very appreciably skew. Therefore, if, for a certain investigation, the measure of dispersion is a more important measure than that of central tendency, no error would ordinarily be made if the mean and standard deviation are chosen, no matter what the reliability of the median may be.

The reader will have noted that measures o reliability are simply measures of dispersion. *Any* measure not infallibly determined may be thought of as one of a population of such measures. It then only remains to calculate a measure of dispersion for this population to secure an index of the reliability of the measure. The measure of dispersion most universally available and most reliable is the standard deviation. The range though frequently available, is very unreliable and should be used for rough or hasty determinations only. The relationship of the five measures of dispersion — standard deviation, mean deviation, 10–90 percentile range, quartile deviation, and the range, to each other will be considered in Section 31 and Problem 1, Chapter V, for the normal distribution, which is probably more typical of uni-modal distributions in general than any other single distribution.

PROBLEMS

1. Calculate the first and second moments from "zero income" for the data of Table X and by proper transformation (*a*) determine μ_2, the second moment from the mean, and (*b*) determine the second moment from the median by formula [28] and check by formula [27].

2. Calculate the standard errors of the (*a*) L. Q., (*b*) Mdn., (*c*) U. Q., (*d*) *M*, for the hypothetical distribution of incomes, Table X. Which is the more accurate average for these data, the mean or the median?

3. Using the grouped data giving changes in wholesale prices, Table XV, determine which is the more reliable average, the mean or the median.

4. (*a*) Which is the more reliable average, the mean or the median, in the case of College Marks, Table XVIII?

(*b*) In this case what is the proper number of class intervals to combine in determining the standard error of the median? [Answer to (*b*): The population, 773, is large and an interval of three units, $\frac{1}{16}$, the range, would be reasonably satisfactory were it not for the fact that there is a decided periodicity, which is irrelevant so far as pupils' talents are concerned, so that the proper interval is one of five units.]

5. (*a*) Determine the standard error of the second moment of the income data, Table X.

(*b*) Determine the standard error of the standard deviation of the same data.

6. Derive μ_3 and μ_4 for frequencies given by the terms of the binomial $(p + q)^n$ in a manner similar to that illustrated for μ_1 and μ_2. Much scratch paper will be needed.

7. Prove that if c is a constant and x a variable then

$$\sigma_{cx} = c\sigma_x.$$

8. Devise a formula similar to [7 *a*] except that the sum of the measures above the mean instead of the sum of those below is involved.

CHAPTER V

THE NORMAL PROBABILITY DISTRIBUTION

Section 25. Derivation of Equation of Normal Distribution

Many frequency distributions are very similar in type. These distributions are characterized by being symmetrical with respect to the mean; by having a single mode which is at the mean: i.e., the slope of the curve at the mean is zero; by tapering off from the mean and in such a manner that the slope again approaches zero as the frequencies or ordinates of the curve approach zero. The symbol y will be used for the ordinate unless $N = 1.0$, in which case z is used to conform with certain tables in this text and with Sheppard's tables. Following Pearson, we may derive the simplest curve which has these characteristics. It is necessary to use the calculus in this derivation, so that one unfamiliar with it may simply note the conclusions.

The differential equation $dy/dx = Cxy$ is an equation, origin at the mean, whose slope is zero both when x is zero and when y is zero. It is the most concise form imposing the required slope conditions of any which has been noted by the writer or any which he is able to conceive. Integrating this equation gives: (All the integration formulas used in this chapter may be found in Peirce, 1910.)

$$y = k\, e^{C\frac{x^2}{2}}$$

If k and C are both positive it is found, by plotting or by more analytical means, that the curve has a minimum instead of a maximum at $x = 0$; also that y does not approach zero for any real value of x. It is therefore necessary that C be negative or setting $C = -c$ the differential equation may be written $dy/dx = -cxy$ and the integral

$$y = ke^{\frac{-cx^2}{2}}$$

Let us investigate the moments of this curve. If N is the total population or total area under the curve

$$N\mu_0 = \int_{-\infty}^{\infty} y\,dx = k\sqrt{\frac{2\pi}{c}} = N$$

$$N\mu_1 = \int_{-\infty}^{\infty} yx\,dx = NM = 0$$

$$N\mu_2 = \int_{-\infty}^{\infty} yx^2\,dx = N\sigma^2 = \frac{k}{c}\sqrt{\frac{2\pi}{c}}$$

Solving the first and third of these equations for c and k gives

$$c = 1/\sigma^2 \text{ and } k = N/\sigma\sqrt{2\pi}$$

This gives as the final equation of the curve

$$y = \frac{N}{\sigma\sqrt{2\pi}}\, e^{\frac{-x^2}{2\sigma^2}} \qquad \text{(The Normal Probability Curve)} \ldots\ldots [44]$$

in which y is the frequency or ordinate corresponding to a deviation x, N is the total frequency, σ the standard deviation of the measures, $\pi = 3.1416$, and $e = 2.7183$ — the Naperian base of logarithms. This equation is identical with the following convergent series:

$$y = \frac{N}{\sigma\sqrt{2\pi}}\left[1 - \left(\frac{x}{\sigma\sqrt{2}}\right)^2 + \frac{1}{2!}\left(\frac{x}{\sigma\sqrt{2}}\right)^4 - \frac{1}{3!}\left(\frac{x}{\sigma\sqrt{2}}\right)^6 + \cdots\right] \ldots [45]$$

Section 26. Certain Properties of the Normal Distribution

The first derivative of equation [44] is:

$$\frac{dy}{dx} = \frac{N}{\sigma\sqrt{2\pi}}\, e^{\frac{-x^2}{2\sigma^2}} - \frac{-x}{\sigma^2} = y = \frac{x}{\sigma^2} \quad \ldots\ldots\ldots\ldots [46]$$

and, as the mode of derivation necessitated, it has a maximum at the mean ($x = 0$) and a zero slope at the extremes ($y = 0$).

The second derivative is:

$$\frac{d^2y}{dx^2} = \frac{-N}{\sigma^3\sqrt{2\pi}}\, e^{\frac{-x^2}{2\sigma^2}}\left(-\frac{x^2}{\sigma^2} + 1\right) \quad \ldots\ldots\ldots\ldots [47]$$

This is zero when x equals plus or minus σ, so that the points of inflection of the normal probability curve are at points one standard deviation above and below the mean.

The first moment, μ_1, for the entire curve is of necessity zero as deviations are measured from the mean, but if the first moment from the mean for half the curve, $\mu_1 \Big]_0^\infty$, is found it will give the average or mean deviation.

$$\mu_1 \Big]_0^\infty = \frac{1}{\frac{N}{2}} \int_0^\infty yx\,dx = \frac{2\sigma}{\sqrt{2\pi}} = .7979\,\sigma \quad \ldots\ldots\ldots [48\,a]$$

It is thus found that the average or mean deviation is .7979 times the standard deviation.

M. Dev., or Av. Dev. $= .7979\,\sigma$ (Relation between average deviation and standard deviation in case of a normal distribution) . . . . . [48]

It is frequently desirable to know how far out, in both directions, it is necessary to go to secure one half the total frequency. This distance is called the probable error because of the fact that if the distribution is one of magnitudes varying by chance from some one magnitude (the mean) then the chances are one to one that any single measure will vary from this magnitude by an amount as great as the probable error.

The area under the curve is given by the integral, $\int z\,dx$. Therefore if the equation

$$\frac{N}{2} = \int_{-x}^{x} y\,dx$$

could be solved for x, it would give that distance which if measured in each direction from the mean would include one half the total population. The integral desired may be expanded into the following convergent series:

$$\int_0^x y\,dx = \frac{N}{\sqrt{\pi}}\left[\frac{x}{\sigma\sqrt{2}} - \frac{1}{3\cdot 1!}\left(\frac{x}{\sigma\sqrt{2}}\right)^3 + \frac{1}{5\cdot 2!}\left(\frac{x}{\sigma\sqrt{2}}\right)^5 - \frac{1}{7\cdot 3!}\left(\frac{x}{\sigma\sqrt{2}}\right)^7 + \cdots\right] \quad \ldots . [49]$$

Setting this equal to $.25\,N$, the number of cases between the mean and plus one probable error, and solving for x gives $.6744898\,\sigma$, the value of the probable error.

Section 27. Kelley–Wood Table of the Normal Probability Integral

The upper limit, x, of the integral, $I = \int_0^x z dx$, when $N = 1$ and $\sigma = 1$, has been evaluated for values of the area, I, by .001's, from .000 to .499 and are tabled in the K–W table,* given in the last pages of this text. The argument for the table is either I, the area from the mean on to the stump of the distribution, q, the area of the smaller portion cut off, or p, the area of the larger portion. I in this table equals $\frac{\alpha}{2}$ of Sheppard's tables, but whereas the tabulated entry in Sheppard's most extensive table is $\frac{\alpha}{2}$ and the argument is x, here the tabulated entry is x and the argument I. In both tables the ordinate is a tabled entry. The two tables supplement each other. Sheppard's tables will be found the more convenient to use if deviates are known and either areas or ordinates desired, while the K–W table will prove the more serviceable if areas are known and deviates or ordinates desired. For expressing a distribution composed of categories arranged in a rank order and having varying frequencies, in terms of a normal distribution, the K–W table is much the more serviceable. Continual reference to Table K–W is made in subsequent chapters of this text and if the meaning of I, q, p, x and z are definitely fixed in mind it will greatly assist in the understanding of subsequent derivations and formulas (cf. pages 371–383).

* The table is called the Kelley–Wood, or K–W, table because Dr. Ben D. Wood calculated by interpolation, using third and fourth order differences, from Sheppard's tables, values of the abscissa x corresponding to areas from $I = .000$ to $I = .400$; because my wife calculated, by formula [49], values at decreasing intervals from $I = .400$ to $I = .499$, and because I calculated by interpolation certain values of the deviate from $I = .400$ to $I = .499$ and also calculated either by interpolation or by the aid of eight place logarithms, values of the ordinate, z. The labor has been substantial and I commend to the inquisitive the calculation of the deviate for $I = .499$, which Mrs. Kelley determined to be equal to 30.9022850+.

Columns I, x and z constitute the basic table of the probability integral, but the added columns z/q, z/p and pq, also calculated by Mrs. Kelley, will be found serviceable in many formulas.

The last figure of the entries in the basic table may be expected occasionally to be in error by 1. — T. L. K.

Section 28. Further Properties of the Normal Distribution

The probable error was found by means of formula [49].

P. E. = .6744898 σ (Probable error of any magnitude in terms of the standard deviation or standard error of the magnitude).....[50]

It is to be noted that the probable error is defined as a certain fixed fraction of the standard deviation, or standard error. The relationship that half the population lies between plus and minus .67449 σ, is strictly true only in case of a normal distribution; however it is the customary measure to use whenever thinking of chance variations, whether the distribution under consideration is normal or not. It must be definitely kept in mind that the P. E. has no status or means of calculation independent of the standard error; it is simply a measure of deviation .67449 times as large as the standard deviation and should not be confused with the quartile deviation which, regardless of the shape of the distribution, is one half the distance from the lower to the upper quartile. From the lower quartile to the upper quartile is always a distance of $2Q$ and is a range that always contains just one half the measures, whereas from 1 P. E. below the mean to 1 P. E. above is a range that contains exactly one half the measures only in the special case when the distribution is normal. It is to be expected that distributions of measures which are composite measures based upon a large number of separate scores will in general more closely approximate a normal distribution than do the distributions of separate scores themselves,* so that the error introduced in thinking of 50 per cent of the cases as lying between + 1 P. E. and − 1 P. E. is very small, if the P. E. under consideration is that of any average, of any coefficient of correlation, of any measure of dispersion, or in fact of any measure whatever derived from a large number of other measures. Quite substantial error may, however, be introduced if the P. E. of the distribution of original measures is taken as such that 50 per cent of the cases lie between

* I have not proven this analytically but have found it to be true with many distributions with which I have had to deal. — T. L. K.

+ 1 P. E. and − 1 P. E. (See problems 2 and 3 at end of chapter.)

Certain important relations between the moments of the normal distribution exist. The third moment, $\mu^3 = \frac{1}{N}\int_{\infty}^{\infty} yx^3 dx$, of course, equals zero as the curve is symmetrical with respect to its origin, the mean.

For the fourth moment we have:

$$\mu_4 = \frac{1}{N}\int_{-\infty}^{\infty} yx^4\, dx = 3\,\sigma^4 \quad \ldots\ldots\ldots\ldots\ldots\ldots [51]$$

These last two relationships are important in that they provide a means of determining how closely given data fit a normal distribution. If $\mu_3 = 0$ and $\mu_4 = 3\,\sigma^4$ the fit is entirely satisfactory and the normal curve will better fit the data than any other uni-modal curve. If these two relationships do not exactly hold, the significance of the discrepancy can be determined by the formulas giving probable errors of any moments, given in the preceding chapter, or more nearly by determining the values and probable errors of two constants β_1 and β_2. These are used in all curve fitting following Pearson's method, and are defined by the equations:

$$\beta_1 = \frac{\mu^2_3}{\mu^3_2} \qquad \beta_2 = \frac{\mu_4}{\mu^2_2} \qquad \text{[Formulas 69 and 70 of Sec. 36]}$$

For a normal distribution $\beta_1 = 0$ and $\beta_2 = 3$. The probable errors of β_1 and β_2 may be found from Tables 37 and 38 of Pearson's Tables. If for any distribution the obtained β's differ from 0 and 3 respectively by amounts which are small with reference to their probable errors the data may be considered normal. The probable errors of these β's will be found to be large if the populations are small. This is simply indicative of the fact that it is impossible to determine the type of a distribution from a small population and it is scarcely worth attempting unless the population is over 100.

Section 29. Properties of Portions of a Normal Distribution

The method followed in the calculation of the average deviation is serviceable in determining the mean deviation of any tail of a normal distribution. Let a "unit normal distribu-

tion" be one of standard deviation and population each equal to 1, then the mean deviation from the mean, of the tail of a normal distribution covering the portion from x to ∞ is given by the equation:

$$\text{M. Dev. of Tail} = \frac{\int_x^\infty yx\,dx}{N_x} = \frac{\sigma^2 y_x}{N_x} = \frac{\sigma z_x}{q_x} \quad \text{(Mean deviation of the tail of a normal distribution)} \ldots\ldots [52]$$

in which y_x is the ordinate per unit base at the point of truncation; N_x is the number of cases lying beyond this point; z_x is the value of the ordinate of a unit normal curve at the stump or point of truncation x, and q_x is the number of cases in the unit normal distribution from the point of truncation x on to ∞. In case of a unit normal distribution we have:

$$\text{M. Dev. of Tail} = \frac{z}{q} \quad \text{(Mean deviation of the tail of a unit normal distribution)} \ldots\ldots\ldots\ldots [53]$$

This magnitude, z/q, is given in Table K–W. In case $q < .5$ use column "z/q" and in case $q > .5$ use column "z/p".

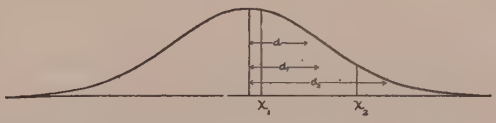
This relationship between ordinate and mean deviation of tail is one of the unique and very interesting properties of the normal distribution. It has many applications, one of which is considered herewith. In case the tail is one half the curve we have: $.7979\,\sigma = \frac{\sigma^2 y_0}{.5N}$, in which y_0 is the ordinate per unit base interval at the mean. Solving for σ gives, approximately,

$$\sigma = \frac{.4\,N}{y_0} \quad \text{(Formula for roughly determining the standard deviation of a distribution which is approximately normal)} \ldots\ldots [54]$$

Accordingly, if a rough estimate of the standard deviation of a distribution will suffice, it may be obtained by dividing .4 of the total population by an estimate of the height of the ordinate, at the mean, of the normal curve which would best fit the data.

A simple extension of the method followed in obtaining the mean deviation of the tail will give the mean deviation from the mean of any part of the distribution. Consider the standard deviation and area of the following figure to be 1 and let it be required to find the mean deviation, from the mean of the entire distribution, of that part of the distribution between x_1 and x_2. Let the ordinates at these points be z_1 and z_2. Let

q_1 and q_2 be the proportions of the population lying above x_1 and x_2 respectively. Let d = the required mean deviation; d_1 = the mean deviation of the tail from x_1 on; d_2 = the mean deviation of the tail from x_2 on. Then $(q_1 - q_2)$ is the pro-



portion lying in the interval from x_1 to x_2. The first moment of the distribution beyond x_1 is equal to the first moment of that part between x_1 and x_2 plus the first moment of that part beyond x_2, or

$$q_1 d_1 = (q_1 - q_2)\, d + q_2 d_2$$

That is, solving

$$z_1 = (q_1 - q_2)\, d + z_2$$

$$d = \frac{z_1 - z_2}{q_1 - q_2} \quad \text{(Mean deviation of a portion of a unit normal distribution)} \ldots . [55]$$

The magnitudes q_1 and q_2 are the proportions lying beyond the upper and lower limits respectively of the class involved, and z_1 and z_2 are the ordinates for these proportions as given in Table K–W.

As an illustration the following problem is given. Assuming a normal distribution, express the following school marks as deviations from the mean:

Marks	Percentage of Pupils Receiving Mark Indicated	q_1	q_2	Calculation from Table K–W		$\frac{z_1 - z_2}{q_1 - q_2}$
				z_1	z_2	
A	11.4	.114	.000	.192900	.000000	1.692
B	34.7	.461	.114	.397034	.192900	.588
C	32.5	.786	.461	.291399	.397034	− .325
D	10.2	.888	.786	.190478	.291399	− .989
E	9.0	.978	.888	.052485	.190478	− 1.533
F	2.2	1.000	.978	.000000	.052485	− 2.386

The table informs us that a mark of A is equivalent to a position 1.692 standard deviations above the mean of the group, that a grade of B is .588 standard deviations above the mean, a grade of C is .325 standard deviations below the mean, etc.

The standard deviation of a portion of a normal distribution is developed in Section 60 in connection with another problem, — see formula [188].

Section 30. The Probability of Exceeding a Given Divergence

The normal curve assists in establishing the degree of confidence which may be placed in statistical findings. The significance of any measure is to be judged by comparison with its probable error. If a child makes a score of 80 on a certain test and if the probable error of the score is 5, we may estimate the chances of the child's true ability being as much as 100. We assume that the distribution of the child's performances would follow a normal curve. Note that the assumption is not that the talents of children in general follow a normal distribution. This latter might be less reasonable than the one we are called upon to make. Moreover, so little difference in probabilities, except for extreme deviates, is ordinarily consequent to differences in forms of distribution, that the assumption of normality is little likely to result in serious error for such problems as the present one. For extreme deviates it generally does not matter so far as any practical deductions are concerned whether the chances are 1 in 1000 or ten times as great. Nor for smaller deviates does it make any particular difference whether the chances are 400 in 1000 or 410 in 1000. Should such differences as mentioned be significant in any particular problem, no assumption should be made, but the type of the curve should be experimentally determined.

For the problem in hand: If the P. E. is 5 the standard error is $\left(\frac{5}{.6745}\right) = 7.413$. The difference between the scores that we are concerned with is (100–80) = 20, which is $\left(\frac{20}{7.413}\right) =$

2.698 standard errors. The K–W Table, or more conveniently for this problem Sheppard's Tables, may be used to find the area in the tail below the point which is 2.698 standard deviations below the mean. The tables give .0035. To interpret this we should postulate the person's true ability as being 100 and his various performances distributing themselves in a normal distribution, with standard deviation equal to 7.413 around this mean. Then .0035 of the area of the curve will lie below the point 80. Accordingly if his true ability is 100, only 35 times in 10000, or 3.5 times in 1000, would a score as low or lower than 80 be expected. With such figures a person could accept the proposition that the child's ability was not as great as 100 with about as much certainty as he can start across a business street expecting not to be hit by an automobile. It is, in other words, just such a conclusion as one is justified in acting upon.

Table K–W is built upon the basis of the standard deviation as the unit of variability, instead of the probable error. If probable errors instead of standard errors are known, the following table may be used for rough work, thus avoiding the labor of division by .6745:

TABLE XXVI

If a difference is x times its probable error	*The Likelihood of a Difference as Great as this Obtained One*									
	and in the same direction, is 100 p in 100, or 100 p chances of its occurring to 100 q chances of its not occurring						and in the same or the opposite direction, is 2 × 100 p in 100, or 200 p chances of its occurring to 100 (1–2 p) chances of its not occurring					
x	100p	in	100	100p	to	100q	200p	in	100	200p	to	100(1–2p)
.5	37	in	100	37	to	63	74	in	100	74	to	26
1.0	25	in	100	25	to	75	50	in	100	50	to	50
1.5	16	in	100	16	to	84	31	in	100	31	to	69
2.0	9	in	100	9	to	91	18	in	100	18	to	82
2.5	5	in	100	5	to	95	9	in	100	9	to	91
3.0	2	in	100	2	to	98	4	in	100	4	to	96
3.5	1	in	100	1	to	99	2	in	100	2	to	98
4.0	.3	in	100.0				.7	in	100.0			
5.0	.02	in	100.00				.04	in	100.00			
6.0	.001	in	100.000				.003	in	100.000			
7.0	.0001	in	100.0000				.0001	in	100.0000			
8.0	.000001	in	100.000000				.000003	in	100.000000			

Section 31. Summary of Facts Concerning the Normal Distribution

A summary of the facts already discovered together with a few determined later in regard to the normal probability curve gives the following:

1. It is uni-modal, symmetrical with respect to the mean, and is completely determined when N, the population, M, the mean, and σ, the standard deviation of the distribution, are known.

2. The mean, median, and mode coincide.

3. Measures of dispersion are related in the following ways:

$$\left.\begin{aligned} Q &= \text{P. E.} = .84535\ \text{A. D.} = .67449\,\sigma = .26315\,D \\ \text{A. D.} &= 1.1829\,Q = 1.1829\ \text{P. E.} = .79788\,\sigma = .31129\,D \\ \sigma &= 1.4826\,Q = 1.4826\ \text{P. E.} = 1.2533\ \text{A. D.} = .39015\,D \\ D &= 3.8001\,Q = 3.8001\ \text{P. E.} = 3.2124\ \text{A. D.} = 2.5631\,\sigma \end{aligned}\right\} \quad \ldots[56]$$

The range covered by the measures is approximately,

In case the total population is	10,	$= 4\,\sigma$
" " " " " "	50,	$= 5\,\sigma$
" " " " " "	200,	$= 6\,\sigma$
" " " " " "	1000,	$= 7\,\sigma$

σ = approximately $.4\,N \div$ the height of the smoothed ordinate at the mean, median, or mode.

4. The points of inflection of the curve are at distances $1\,\sigma$ and $-1\,\sigma$ from the mean.

5. Every odd moment $\mu_1, \mu_3, \mu_5 \cdots$ of the curve is equal to 0. The even moments are given by

$$\left.\begin{aligned} \mu_2 &= \sigma^2 \\ \mu_4 &= 3\,\mu^2{}_2 = 3\,\sigma^4 \\ \mu_6 &= 15\,\mu^3{}_2 = 15\,\sigma^6 \\ \mu_8 &= 105\,\mu^4{}_2 = 105\,\sigma^8 \\ \beta_1 &= 0, \text{ and } \beta_2 = 3. \end{aligned}\right\} \quad \ldots\ldots\ldots\ldots[57]$$

6. The mean deviation of a truncated portion of the curve, taken from the mean of the entire distribution, is equal to the square of the standard deviation of the entire distribution into the height of the curve at the point of truncation, divided by the number of cases in the tail.

7. The most reliable constant of the distribution is the standard deviation. Its probable error =

$$\frac{.6744898\,\sigma}{\sqrt{2N}}, \text{ or } \frac{.477}{\sqrt{N}} \text{ of its own magnitude} \ldots\ldots\ldots [58]$$

This follows from formulas [32-a] and [50].

The probable error of the average deviation =

$$\frac{.4066\,\sigma}{\sqrt{N}}, \text{ or } \frac{.510}{\sqrt{N}} \text{ of its own magnitude} \ldots\ldots\ldots [59]$$

The probable error of D, the 10–90 percentile range, =

$$\frac{1.5373\,\sigma}{\sqrt{N}}, \text{ or } \frac{.600}{\sqrt{N}} \text{ of its own magnitude} \ldots\ldots\ldots [16\,a]$$

The probable error of the quartile =

$$\frac{.5306\,\sigma}{\sqrt{N}}, \text{ or } \frac{.787}{\sqrt{N}} \text{ of its own magnitude} \ldots\ldots\ldots [60]$$

This follows from formulas [14] and [50].

It is thus seen that if N measures result in a certain reliability in the standard deviation, it requires to obtain an equal reliability, $1.14\,N$ measures in the average deviation, $1.58\,N$ measures in the 10–90 percentile range, and $2.72\,N$ measures in the quartile deviation.

8. Measures of central tendency are less reliable than measures of dispersion. Little, if any, significance attaches to a measure of the unreliability of an average expressed in terms of itself, and, furthermore, since in the normal distribution all measures of central tendency coincide, it will suffice for purposes of comparison to give the probable error of each.

$$\text{P. E. of mean} = \frac{.6745\,\sigma}{\sqrt{N}} \quad \text{(Normal or any other distribution)} \ldots\ldots\ldots [61]$$

$$\text{P. E. of median} = \frac{.84535\,\sigma}{\sqrt{N}} \quad \text{(In case of normal distribution only)} \ldots\ldots\ldots [62]$$

P. E. of the mode is unknown unless the mode is determined from the equation which best fits the data, in which case its probable error compares favorably with those of the mean and median.

It is seen that if N measures result in a certain reliability in the mean, it requires $1.57\,N$ measures to obtain an equal reliability in the median.

9. If a distribution is normal the most reliable measure of dispersion based upon percentiles is that between the 7th and 93d percentiles. Of almost as great reliability is the 10–90 percentile range.

10. The distributions of frequencies in the point binomial $(p + q)^n$ closely approximates a normal distribution if n is large and neither p nor q very small. For n infinite and neither p nor q infinitesimal the point binomial distribution becomes a point normal distribution.

11. The average deviation from the mean of any portion of a normal distribution may be obtained from the equation:

$$d = \frac{z_1 - z_2}{q_1 - q_2},$$

in which the q's are proportions of the population and the z's are corresponding ordinates as given in Table K–W.

12. The standard deviation from the mean of any portion of a normal distribution may be obtained from the equation:

$$\sigma^2_1 = 1 + \frac{x_1 z_1 - x_2 z_2}{q_1 - q_2} - d^2 \quad \text{[Section 63, Formula 188]}$$

13. The equation of the normal distribution is

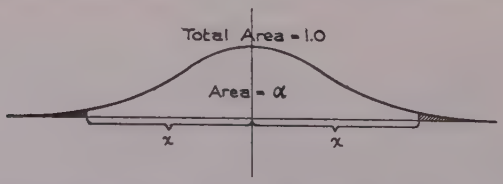
$$z = \frac{N}{\sigma\sqrt{2\pi}} e^{\frac{-x^2}{2\sigma^2}} \quad \text{.....[44]}$$

or,

$$z = \frac{N}{\sigma\sqrt{2\pi}}\left[1 - \left(\frac{x}{\sigma\sqrt{2}}\right)^2 + \frac{1}{2}\left(\frac{x}{\sigma\sqrt{2}}\right)^4 - \frac{1}{3!}\left(\frac{x}{\sigma\sqrt{2}}\right)^6 + \cdots\right] \text{..[45]}$$

PROBLEMS

1. Given a normal distribution with areas and deviations as indicated in the accompanying figure, then $(1 - \alpha)/1$ is the probability of a measure



lying in the shaded portion or, in other words, of a measure deviating from the mean by a distance greater than x. If the probability of a single

measure lying beyond x is this small amount $1 - \alpha$, then the probability of a measure, in case of a population of N measures, lying beyond this point is $N(1 - \alpha)$. If this probability, $N(1 - \alpha)$, equals .5, then the value x corresponding to the α is such a deviation that the chances that a measure will lie beyond the point x is just equal to the chance that no measure will lie beyond it. The distance x is therefore the most probable maximum deviation which will be found in the case of a population of N. As a sufficiently close approximation x may be taken as equal to one half the range. Accordingly using Table K–W the following table is obtained:

Range	x	$1 - \alpha$	N, Such That $N(1 - \alpha) = .5$
3σ	1.5σ	.1336	
4σ	$2.\ \sigma$	.0455	9
5σ	2.5σ	.01242	40
6σ	$3.\ \sigma$	.00270	185
7σ	3.5σ	.000465	1075
8σ	$4.\ \sigma$	.0000634	

Complete the table, determining values for 3σ and 8σ. [Answer: If the population is 4 (more exactly 3.75) the range of the measures is (providing the total distribution from which the sample of 4 is drawn is normal) most probably equal to 3σ; and if the population is 8660 the range is most probably equal to 8σ.]

2. In the case of the distribution of incomes given in Table X calculate the L. Q. and the U. Q. and the points corresponding to − P. E. and + P. E. Compare values found. What percentage of the cases lie between these + and − P. E. points?

3. Do the same for the distribution of Wholesale Price Indexes given in Table XIV.

4. Estimate the standard deviation of the distribution of temperatures given in Table VIII and Charts I and II by first estimating the height at the mean of the normal curve which would seem to fit the data.

5. Do the same for the College Marks data given in Table XVIII. Compare σ found with the correct σ.

6. Group the College Marks data in fives, 47–52 constituting one group, 52–57, the next, etc. Plot and from height of the curve at the mean, estimate the σ. Compare with correct value. What adjustment in estimating σ by this short method is necessary in case the data are grouped? [Answer: The obtained σ is in terms of intervals and must be multiplied by the number of elementary units in each group to give the σ expressed in elementary units.]

7. Verify the calculation of equivalent scores given in Table XXXV.

8. If the plumage of certain fowl is either blue, splashed, or white, and if the percentages in these categories are 28, 60, and 12, what numerical values should be assigned to these colorations should it be desired to treat them as color deviations in a normal distribution?

9. Assuming normality of distribution in the temperature data, Table VIII, and using 81.548 and 6.190, the values of the mean and standard deviation already found, calculate the ordinate at + 1 P. E., 85.723, and compare with the actual ordinate. [Answer: Theoretical 3.17, Actual without smoothing 3.00.] Still assuming normality, what is the average deviation from the mean of the truncated portion beyond this point? [Answer: 7.86.] Of the portion below this point? [Answer: −2.62.]

10. Verify all statements in paragraph 7, Section 31.

11. Verify statement in last sentence of paragraph 8, Section 31.

12. (*a*) Calculate β_1 and β_2 for the point binomial when $p = q = 1/2$ and $n = 25$. [Answer: $\beta_1 = 0$, $\beta_2 = 2.92$.]

(*b*) Calculate β_1 and β_2 for the point binomial when $p = .1$, $q = .9$ and $n = 25$. [Answer: $\beta_1 = .2844$, $\beta_2 = 3.204$.]

(*c*) Calculate β_1 and β_2 for the point binomial when p and q are both finite and $n = \infty$. [Answer: $\beta_1 = 0$, $\beta_2 = 3$.]

CHAPTER VI

COMPARABLE MEASURES

Section 32. The Conditions Requisite for Comparison

In many studies measures of the same, or nearly the same, phenomena are obtained and it is desired to compare results. Gross measures or scores can with validity be compared directly only in case they are in the same units and have been obtained under very similar conditions. There are four methods in common use, the purpose of each of which is to derive comparable measures from original scores obtained in such manner as not to be directly comparable. Of these four the first and the only one which is universally sound is that based upon the complete equivalence of the scales of measurement involved; a second is the ratio or index method; a third may be called the equivalence of standard measures method; and a fourth may be called the equivalence of successive percentiles method.

The first method presupposes that the complete equivalence between measures is known. If both are rectilinear scales and two points of the one have been determined to be equivalent to two points of the other, then for every point of the one an equivalent point on the other may be immediately located. As an illustration of this method may be considered the comparison of two heights, one expressed in centimeters and the other in inches. In the case of inches and centimeters the two points which have been determined as equal are:

0.0 centimeter = 0.0 inch
100.0 centimeters = 39.37 inches

This type of equating is common both in the physical sciences and in the social sciences, but it should be noted that it is entirely sound only in case the two scales measure identically

the same thing in the same linear manner. Any number of functions may be found which agree at two or more points, but are not identical, such, for example, as, $f' = \sin^2 x; f'' = \frac{2x}{\pi}$; etc. For each of these the function equals zero when x equals zero and the function equals 1 when x equals $\pi/2$, but in general $f' \neq f''$.

The minimum number of conditions which must be met before two scales can be fully equated are three. The conditions are, (*a*) one point of the first must be known to be equal to a point of the second, (*b*) a second point of the first must be known to be equal to a second point of the second, and (*c*) the law establishing the relationship between successive points on the first must be known to be the law underlying the second. This third condition is the hardest to establish and should be examined the most critically. Even in the physical sciences it frequently can only be approximately established. Compare, for example, the relation between temperature, pressure and volume in the case of two gases. When these three conditions are met the determining of equivalent scores is simple and is just such a problem as that of finding equivalent temperatures in the centigrade scale to those in the Fahrenheit scale, knowing that 0° and 100° centigrade correspond to 32° and 212° Fahrenheit respectively and that both scales are rectilinear.

It frequently happens that only two of the three conditions mentioned are established, in which case a guess is sometimes made as to the third and an equating attempted. The excellence of the resulting system of equivalent measures is uncertain, and all interpretations drawn should be with the reservation that they are subject to the validity of the assumption involved.

Section 33. The Ratio Method

In case conditions (*a*) and (*b*) are met, and condition (*a*) is "a score of zero on the one scale is equal to a score of zero on the other," condition (*c*) is frequently assumed to be "the same proportion between the units of the two scales maintains throughout." With these underlying conditions the ratio

method is frequently used. Illustrations will show the hazards involved. Given the following sets of data:

TABLE XXVII

	Height in Cm.	Weight in Lbs.
Individual A	138	75
Average adult	172	145

(Data for individual A are those given in Whipple for the average 12.0 year old boy.)

TABLE XXVIII

Weight			
Elephant A	4000 pounds	Butterfly B	2 grams
Average for species	3600 pounds	Average for species	1 gram

TABLE XXIX

United States Bureau of Labor Statistics — Average Aug. 15 Retail Prices

1918	Fresh Eggs 53.6¢ doz.	Potatoes 3.9¢ pd.	Bread 9.9¢ pd.	Tea 65.8¢ pd.
Average 1913–17	35.8	2.2	7.3	55.7

If one is attempting to secure a maturity measure based upon height and another based upon weight one might start with the following propositions:

(*a*) 0 cm. height indicates the same mount of maturity as 0 pounds weight, (*b*) 172 cm. height indicates the same amount of maturity as 145 pounds weight, (*c*) the law of development of height is the same as that for weight. Of these three statements (*a*) is probable entirely sound, (*b*) probably tolerably satisfactory, particularly if dealing with groups and averages, while (*c*) is probable quite absurd. Accepting these three propositions is equivalent to saying that scores X_1 and X_2 in the two measures, which satisfy the following equation, in which M_1 and M_2 are the means of the two series, are equivalent:

$$\frac{X_1}{M_1} = \frac{X_2}{M_2}$$

The ratio is often used with some other magnitude than the mean as a base so that a more general statement of the equation connecting equivalent scores is:

$$\frac{X_1}{B_1} = \frac{X_2}{B_2} \quad \text{(Equivalent scores upon the assumption of equality of ratios)} \ldots\ldots\ldots\ldots\ldots [63]$$

B_1 and B_2 should be values of the variables which are known with more than usual certainty to be comparable and reliable. It is also desirable that they be not small with reference to the scores involved. Due to the greater reliability of means than of individual scores the use of the mean as a base has much to recommend it. Letting σ_1 and σ_2 stand for the standard deviations of the X_1 and X_2 scores, one criterion of the soundness of the assumption of the equality of ratios is:

$$\frac{B_1}{\sigma_1} = \frac{B_2}{\sigma_2} \quad \text{(Criterion to use in judging of the appropriateness of the ratio method)} \ldots\ldots [64]$$

The use of this criterion is illustrated in the next section in a problem in which the bases are the means.

The calculated ratio scores of Individual A are not equal, for A stands ($138/172 =$) .802 on the height maturity scale and ($75/145 =$) .517 on the weight maturity scale. Accepting proposition (*c*) one would conclude that individual A is a very abnormal person, being some 28.5 per cent more developed in height than in weight. In dealing with mental traits not amenable to direct observation a conclusion equally as absurd as that just drawn might pass for years without discovery. In the case of height and weight the fallacy can be immediately detected and a method followed which will be more reasonable, though it is impossible to say that it is entirely sound, as the proposition (*c*) is still an assumption.

Height being a one-dimensional magnitude and weight approximately three-dimensional (*a*) and (*b*) stand as before and the third becomes: (*c*) The law of development of height is the same as that for the cube root of weight. The comparisons then are: Maturity index based upon height = .803. Maturity index based upon weight $= \sqrt[3]{75/145} = .803$. Upon the basis of these two figures one would conclude that the individual is equally developed in the two traits. This illustration is given to show the material differences which result from

different assumptions as to the laws connecting successive scores of two scales and not to suggest that either of the two methods followed is established as sound. At best, in the problem in question, propositions (*b*) and (*c*) are questionable. Logically proposition (*a*) seems sound, but there are many situations in psychology and economics where a similar statement would be very fallacious.

The hazards of the ratio method are not lessened when dealing with the same sort of function of different things. For example, the weight of one child expressed as a proportion of the average adult weight in comparison with the weight of a second similarly expressed may be very misleading. The two children may have very different hereditary endowments, the one becoming a normal adult of weight 120 pounds and the other a normal adult of weight 145 pounds. The fallacy in using indexes in the case just mentioned is the same as that for Table XXVIII. Elephant A has a weight index of 1.11 and Butterfly B one of 2.00. This constitutes no proof that as a butterfly B is more exceptional than is A as an elephant. It might be true that 10 per cent of butterflies exceed 3 grams in weight and but 5 per cent of elephants exceed 4000 pounds. The indexes do not tell us, but in such case it would seem reasonable to call A the more exceptional.

Using the Labor Bureau data of Table XXIX we find that the 1918 August 15 price of fresh eggs is 150 per cent of the average August 15 price for the years 1913–17; of potatoes 177 per cent; of bread 136 per cent; and of tea 118 per cent. These four ratios tell an important story, but at the same time they may be misleading and for the same reason that the weight ratios of elephants and butterflies are misleading. The law covering the fluctuation of potato prices is almost certainly different from that covering the fluctuation of bread prices and similarly for any two of the products which may be compared. Conditions (*a*) and (*b*) may be fairly sound, but very questionably so of condition (*c*):

(*a*) 0 ¢ per dozen eggs indicates the same sort of a price condition as 0 ¢ per pound for potatoes.

(*b*) 35.8 ¢ per doz. eggs indicates the same sort of a price condition as 2.2 ¢ per pound for potatoes.

(*c*) The conditions determining the fluctuations in the prices of eggs are proportional to those determining fluctuations in potato prices.

Because of the peculiar difficulty of establishing condition (*c*) the ratio method for economic and psychological problems may be expected to be an artifact and not an exact quantitative procedure.

A part of the error involved in combining price ratios of separate items to obtain a general index may be eliminated by weighting the separate ratios inversely as the squares of their variabilities, as proven in Section 91 and illustrated in Section 90. This method, however, will not result in as great accuracy as will one based upon the multiple correlation and regression of the prices involved. Further considerations are given in Chapter XIII.

Section 34. The Standard Measure Method

This is an outgrowth of the method used by Francis Galton. It has certain refinements in the measures involved, but rests upon practically the same principle. Galton considered two measures which attempted to measure the same function to be comparable when each was expressed as a deviation from the median of the group to which it belonged and when each such deviation was divided by the quartile deviation of the group. The three propositions essential to the soundness of this procedure are:

(*a*) The median score of the first measure indicates the same sort of a condition as the median score of the second measure.

(*b*) A score of the first measure which deviates one quartile from the median indicates the same sort of a condition as a score of the second which deviates in the same direction one quartile from its median.

(*c*) In general, deviations of the two measures which are in the same proportion as the quartile deviations are indicative of the same sort of a condition.

More briefly stated these propositions are.

(*a*) Median scores are comparable.

(*b*) Quartile deviations are comparable.

(*c*) The same proportions as between quartiles holds for all equivalent deviations from the medians.

Since the mean can generally be more reliably determined than the median, and the standard deviation than the quartile deviation, the Galton procedure has been dropped and the following propositions taken as a basis:

(*a*) Mean scores are comparable.

(*b*) Standard deviations are comparable.

(*c*) The same proportion as between standard deviations holds for all equivalent deviations from the mean.

Let

$$z_1 = \frac{X_1 - M_1}{\sigma_1}, \text{ and } z_2 = \frac{X_2 - M_2}{\sigma_2} \quad \text{(Standard measures)} \ldots [65]$$

Then the measures to be compared are z_1 and z_2. Such measures as these may be called "standard measures" as they are measures of deviation expressed in terms of standard deviations. The last proposition may then be stated:

(*c*) Equal standard measures are comparable.

It should be noted that there is no implication that a zero score in the first measure is equal to a zero score in the second measure. Proposition (*c*) always needs experimental verification, but for the usual distributions found in the social sciences it seems reasonable to expect that if the means of the distributions are set equal, and if points one standard deviation away from the respective means be placed together, a better approximation to complete equivalence throughout the entire scales will be obtained than if the means and zero points are equated and other values taken in proportion. The following data taken from Pintner (1914) and Kelley (1914 comp.) * will illustrate the method and they also are such as do not

* A numerical error occurs in this reference, the figures herewith presented being the correct ones.

reveal without statistical analysis the inaccuracy of the ratio method:

TABLE XXX

No. of Sample	Mean Scores Given to Samples of Handwriting Upon	
	Ayres Scale	Thorndike Scale
12	20.6	5.9
6	24.2	6.5
8	28.4	7.3
21	35.3	8.4
4	36.2	8.0
15	36.3	8.3
1	37.1	8.1
22	40.3	8.9
5	40.3	9.0
17	41.8	8.9
18	48.9	10.1
14	49.2	10.2
9	52.4	10.7
7	55.7	10.6
24	55.7	10.8
11	56.0	10.7
10	56.9	11.3
2	57.7	10.9
13	58.0	11.2
19	58.9	11.5
20	64.2	11.8
23	74.2	13.8
3	80.1	14.2
16	82.1	14.8

Calling the Ayres X_1 scores and the Thorndike X_2 scores and calculating the required constants yields:

$M_1 = 49.60 \quad \sigma_1 = 15.93 \quad M_2 = 10.08 \quad \sigma_2 = 2.229$

X_1's and X_2's satisfying the following equation are comparable measures:

$$\frac{X_1 - M_1}{\sigma_1} = \frac{X_2 - M_2}{\sigma_2}$$ (Equivalent scores upon the assumption of equality of standard measures) . . . [66]

Solving for certain values yields the equivalent scores given in the first two columns of the following table, XXXI. Treating the same data by the index method gives the equation:

$$\frac{X_1}{49.60} = \frac{X_2}{10.08}$$

Scores which are equivalent as derived from this equation are given in the last two columns of the table.

TABLE XXXI

Standard Measures Method		*Ratio Method*	
Equivalent Scores		Equivalent Scores	
Ayres	Thorndike	Ayres	Thorndike
X_1	X_2	X_1	X_2
− 22.4	0.0	0.0	0.0
0.0	3.1		
20.5	6.0	29.5	6.0
49.6	10.1	49.6	10.1
70.0	12.9	70.0	14.2
84.8	15.0	73.8	15.0

The two methods lead to different results and a very brief study of the original data shows that the equivalents obtained by the standard measure method are much the more reasonable. The fundamental error in this problem of the ratio method is in the assumption of equality of zero scores. That this is an error would not be self-evident to the user of the scales, as samples of handwriting of less merit than 20 on the Ayres scale or 6.0 on the Thorndike are seldom found, so that what constitutes a sample of zero merit on either scale is quite unknown. A similar observation applies to economic situations, for who has experience with, or knows the meaning of, o ¢ as the cost of, let us say, a pound of bread?

Reference to the equations giving equivalent scores shows that knowledge of the means, in case the means are the bases, is all that is necessary to determine the equation giving equivalent scores in the case of the ratio method; but that an added item of information, the standard deviations, is required in the case of the standard measure method. If equivalent measures really are proportionate as assumed by the index method, the equating of standard measures results in the same set of equivalents as given by the ratio method. This special case exists when

$$\frac{M_1}{\sigma_1} = \frac{M^2}{\sigma^2}, \text{ for then } \frac{X_1 - M_1}{\sigma_1} = \frac{X_2 - M_2}{\sigma_2} \text{ reduces to } \frac{X_1}{M_1} = \frac{X_2}{M_2}$$

Accordingly the standard measure method is the more general and contains the ratio method as one of its special cases.

Section 35. The Equivalence of Successive Percentiles Method

This method involves no assumption that the law covering the relation between successive scores is of any particular type other than that involved in the statement "the larger the score the greater the trait, or characteristic, being measured!" Otis (1916) and (1918) in dealing with paired measures, has used a graphic method which gives a line of "rank relation." His method, equivalent to setting the lowest score in series one equal to the lowest score in series two, the next lowest in series one equal to the next lowest in series two, etc., could be called "the equivalence of successive ranks" method, but the title here given is used as being the more general. The method does not depend upon paired measures or upon having two series of the same population, though if measures are paired and high correlation exists between them the reliability of equatings is greatly increased.

Letting P stand for percentiles in the first series and P' for those in the second, the method assumes that equivalent scores are $P_{.01}$ and $P'_{.01}$; $P_{.02}$ and $P'_{.02}$; etc.; and in general

P_p is equivalent to P'_p (Comparable percentiles) . . . .[67]

No single one of these equivalents $P_{.01} = P_{.01}$, etc., can be determined with the reliability that appertains to $M = M'$, or $\sigma = \sigma'$, but, unless it has been experimentally determined that relationships between the two series are rectilinear, or curvilinear according to a known law, a more accurate total set of equivalents may be expected from this method than from either of the two preceding. Objections to the method are, first, that no concise algebraic statement of relationship comes from it and second, that it is responsive to chance oddities in distributions. This second objection can be largely overcome by smoothing graphically as does Otis or by a moving average, as will be illustrated, using the data upon handwriting.

There are but 24 samples of handwriting so that a percentile below the 4.1667th cannot be calculated except by an arbitrary assumption as to what constitutes the lower limit of the interval corresponding to the lowest score. We will therefore begin with the 5th percentile and, to shorten the work, proceed by fives to the 95th.

TABLE XXXII

Percentiles	Equivalent Handwriting Scores		Smoothed Equivalent Scores	
	Ayres Scale	Thorndike Scale	Ayres	Thorndike
5	23.18	6.34	23.1	6.35
10	28.52	7.20	26.7	7.30
15	34.19	7.89	32.4	7.90
20	36.15	8.17	34.2	8.20
25	36.70	8.35	36.0	8.50
30	38.935	8.68	37.8	8.80
35	40.145	8.86	39.6	9.10
40	43.65	9.31	42.3	9.40
45	48.31	10.03	46.8	9.80
50	50.80	10.40	49.5	10.25
55	54.23	10.66	54.15	10.45
60	55.31	10.72	55.05	10.65
65	56.21	10.81	55.95	10.85
70	57.13	11.01	56.85	11.05
75	57.85	11.25	57.75	11.25
80	59.07	11.45	59.1	11.55
85	64.61	12.11	64.5	12.25
90	73.97	13.52	74.4	13.45
95	80.31	14.40	79.8	14.35

TABLE XXXIII

Differences between Successive Five-Percentiles

Raw Percentiles		Smoothed Percentiles	
Ayres	Thorndike	Ayres	Thorndike
5.34	.86	3.6	.95
5.67	.69	5.7	.6
1.96	.28	1.8	.3
.55	.18	1.8	.3
2.235	.33	1.8	.3
1.21	.18	1.8	.3
3.505	.45	2.7	.3
4.66	.72	4.5	.4
2.49	.37	2.7	.45
3.43	.26	4.65	.2
1.08	.06	.9	.2
.90	.09	.9	.2
.92	.20	.9	.2
.72	.24	.9	.2
1.22	.20	1.35	.3
5.54	.66	5.4	.7
9.36	1.41	9.9	1.2
6.34	.88	5.4	.9

The smoothed percentile scores have been calculated from the original series after grouping the Ayres data in 3's (score 21, frequency 1; sc 24, f 1; sc 27, f 1; sc 30, f 0; sc 33, f 0; sc 36, f 4, etc.) and the Thorndike scores in 5's (sc 60, f 1; sc 6.5, f 1, etc.) A moving average would probably lead to slightly better results, but would be laborious with the uneven spacing here present in the scores.

We may judge of the excellence of the two sets of equivalent scores, since the drawing up of a correlation table for the data of Table XXX shows that the relationship between the two scales is almost exactly rectilinear, so that differences between the percentiles upon the one scale should be proportionate to the differences upon the other scale. Columns 1 and 2 of Table XXXIII give these differences for the raw data and columns 3 and 4 give the differences determined from the smoothed data. Rather better results are obtained from the raw data than from the grouped, as would be expected from data showing the high degree of correlation here present. The small fluctuations are, in material part, not random, but genuine, and the grouping process has therefore distorted the facts.

This method of equating scores is thoroughly empirical and therefore applicable to situations in which the law of relationship between variables is unknown, or at least cannot be stated in a simple algebraic formula, but in which sufficient reason exists to warrant the equating.

If several series are to be equated a very serviceable modification of the preceding method is to equate each series, not to any one of them, but to a normal distribution. This can be done, using formula [55], giving by the aid of Table K–W the mean deviation of a portion of a normal distribution. An illustration will make clear the steps involved:

It is frequently desired to compare the performances of pupils receiving marks in different subjects. If the pupils have no subjects and no teachers in common, this can only be done by making some assumption. If there are three teachers, each with 50 pupils, it is more reasonable to assume that the mean abilities of the three groups are equal than that similar literal or percentage grades of the three teachers are equivalent. The data of Table XXXIV present the problem.

TABLE XXXIV

Marks Used by First Teacher	Percentage Given Mark Indicated	Marks Used by Second Teacher	Percentage Given Mark Indicated	Marks Used by Third Teacher	Percentage Given Mark Indicated
A	2.0	A +	.7	1	4.3
B	17.1	A	13.9	2	37.7
C	31.3	A −	4.5	3	50.3
D	40.0	B +	4.6	4	7.7
E	7.7	B	29.4		
F	1.9	B −	4.3		
		C +	4.7		
		C	22.7		
		D	9.2		
		E	6.0		

It is obvious that a mark of A given by the first teacher indicates greater merit than a mark of A given by the second teacher. Equating each mark to a standard-measure score in a normal distribution gives:

TABLE XXXV

Marks Used by First Teacher	Equivalent Standard Measure	Marks Used by Second Teacher	Equivalent Standard Measure	Marks Used by Third Teacher	Equivalent Standard Measure
A	2.4	A +	2.8	1	2.1
B	1.3	A	1.5	2	.8
C	.4	A −	1.0	3	−.5
D	− .6	B +	.8	4	− 1.9
E	− 1.6	B	.3		
F	− 2.5	B −	− .1		
		C +	− .2		
		C	− .6		
		D	− 1.3		
		E	− 2.0		

The method requires little time, but were such equatings being done for a large number of classes a still briefer method could be followed. Instead of finding the mean standard deviation score for the upper 2 per cent, we may find the median: 1/2 the percentage of A's = 1.0, therefore from Table K–W 2.3 is the standard deviation score which is equivalent to the mark of A given by the first teacher. The percentage of A's plus $\frac{1}{2}$ the percentage of B's = 10.6, therefore 1.2 is the score equivalent to B. Similarly, .4 is equivalent to C; − .5 to D; − 1.6 to E; and − 2.4 to F.

The marks given by the second teacher are typically those of a careful grader and show more discrimination than do those of either the first or third teacher, but nevertheless it is more reasonable to assume a normal distribution of talent than such a tri-modal distribution as is indicated by the second teacher's marks. The method may frequently be used for the single purpose of warping data showing an extreme distribution into a more reasonable mold.

The observation has been made that in order to be comparable the two series should be independent measures of the same thing. It is shown in Section 56 how certain correlation functions enable one to estimate whether two series of scores are measures of the same thing. In general it is not necessary that a raw correlation between the two series approaching 1.00 be found, but merely that a coefficient of correlation corrected for attenuation of 1.00 be present.

CHAPTER VII

THE FITTING OF CURVES TO DISTRIBUTIONS

Section 36. Methods of Fitting Curves to Observations

The properties of the normal distribution as given in Chapter V are such that if data fall approximately into this form their interpretation and treatment are frequently greatly simplified. As a practical matter it is often serviceable to treat data as normal even though slight divergence from normality may be known to exist. Probably, however, the majority of distributions cannot by any stretch of interpretation be considered normal. In such case one may resort to one of two procedures, (*a*) either warp data into a normal mold by transformation devices, or (*b*) discard the concept of normality altogether and endeavor to discover an equation which does describe the data.

The equation of the normal curve is

$$y = \frac{N}{\sigma\sqrt{2\pi}} e^{\frac{-x^2}{2\sigma^2}}$$

Not counting N, the population, which does not affect the type of curve, there is only one degree of freedom in this curve since σ is the only constant which is to be determined from the data. To permit of greater freedom one could start as did Edgeworth with an equation of the tvpe

$$y = \frac{N}{\sigma_f\sqrt{2\pi}} e^{\frac{-f^2}{2\sigma^2 f}}$$

in which f is some function of x. As $f(x)$ is made more and more general, greater and greater freedom is given. Other variations of this approach have been followed by Edgeworth (1904), Kapteyn (1903), Thiele (1903) and Charlier (1906). Pearson has criticized this method because the function built

up is what he terms a "shadow function," something not corresponding to any physical measurement, not representing any relationship which is in itself capable of independent interpretation; and as a procedure which tends to make a fetish of the normal distribution. However, should this ghost take on flesh and bone and be found, in certain important cases, to be a measure of what would seem to be a causal force, the method would be amply justified. Judgment may well be held in abeyance pending further experimental treatment. Later in this chapter the normal distribution will be shown to hold a unique and peculiarly dominant position among all the Pearson curves, but this is not an argument for arbitrarily forcing data into this form. It is rather an argument for the study of the features of a given distribution which diverges from this form. The first four sections of this chapter are concerned with the practical details of curve fitting while the theme of the last two sections is the bearing of types of distributions upon problems of stability and trends in evolution.

Section 37. The Principle Underlying Pearson's Method of Curve Fitting

Pearson imposes certain very broad conditions upon the differential equation of the curve. These conditions are so general that many varieties of non-bi-modal distributions are represented. These include (*a*) curves with a maximum frequency somewhere between the limits of the range, called "i-shaped" curves, (*b*) such as have an anti-mode, or point of minimum frequency between the limits of the range, called "u-shaped" curves, and (*c*) such as have no mode, called "j-shaped" curves. The present treatment will describe the calculation of a few of the more important of the fifteen Pearson types, and will present such criteria as are necessary in determining the type of curve to which given data belong, so that one may then go to Pearson's Tables (1914 tables) and other sources, Elderton (1906), Pearson (1894), (1890 and sup 1901), (1902 sys), (1906 skew), (1915 cert) and (1916 app), and determine the equation of the curves.

The fundamental proposition in Pearson's method is that in order to have a good fit the first four moments of the data

should equal the first four moments of the derived equation and second that formula [81] expresses the general differential equation covering all uni-modal curves. The moments are fundamental and may be obtained by aid of the accompanying formulas.

Let the required moments be μ_1, μ_2, μ_3, μ_4.

Let the four moments from the mean, but uncorrected for grouping be ν_1, ν_2, ν_3, ν_4.

Let the raw moments from the arbitrary origin be $\bar{\nu}_1$, $\bar{\nu}_2$, $\bar{\nu}_3$, $\bar{\nu}_4$. Then the following equations lead to the calculation of the μ's:

$$\bar{\nu}_1 = \frac{\Sigma X}{N}, \bar{\nu}_2 = \frac{\Sigma X^2}{N}, \bar{\nu}_3 = \frac{\Sigma X^3}{N}, \bar{\nu}_4 = \frac{\Sigma X^4}{N}$$

$\nu_1 = \bar{\nu}_1 - \bar{\nu}_1 = 0$	(Moments from the mean, knowing them from an arbitrary origin).....see	[24] [21]
$\nu_2 = \bar{\nu}_2 - \bar{\nu}^2{}_1$		
$\nu_3 = \bar{\nu}_3 - 3\,\bar{\nu}_2\bar{\nu}_1 + 2\,\bar{\nu}^3{}_1$		
$\nu_4 = \bar{\nu}_4 - 4\,\bar{\nu}_3\bar{\nu}_1 + 6\,\bar{\nu}_2\bar{\nu}^2{}_1 - 3\,\bar{\nu}^4{}_1$		

Continuing

$\mu_1 = \nu_1 = 0$	(Sheppard's corrections applied to moments from the mean)	[68]
$\mu_2 = \nu_2 - \frac{1}{12}$		[68 *a*, see also Sec. 47]
$\mu_3 = \nu_3$		[68 *b*]
$\mu_4 = \nu_4 - \frac{\nu_2}{2} + \frac{7}{240}$		[68 *c*]

Sheppard's corrections are for an error in the moments due to grouping. They are to be used in case of "high contact"; that is, when the curve approaches asymptotically the base line, or x-axis, at both extremities. In case high contact at both extremities is not present, corrections as given by Pairman and Pearson (1919) should be used.

It should be noticed that the $\bar{\nu}$'s are here defined as were the $\bar{\mu}$'s in Section 21, that the ν's here are the same as the μ's in that section, and that the μ's here differ slightly from the ν's (or the μ's of Section 21), being corrected for a grouping error.

Certain derived constants, β_1, β_2 and the criterion κ_2 are also needed in determining the type to which given data belong. In earlier work in curve fitting a criterion κ_1 was used and though it is not as general a criterion as κ_2 it has much theoretical interest.

$$\beta_1 = \frac{\mu^2_3}{\mu^3_2}, \qquad \text{(One measure of skewness)} \ldots [69]$$

$$\beta_2 = \frac{\mu_4}{\mu^2_2} \qquad \text{(One measure of kurtosis)} \ldots [70]$$

$$\kappa_1 = 2\,\beta_2 - 3\,\beta_1 - 6 \qquad \text{(Criterion } \kappa_1) \ldots\ldots\ldots [71]$$

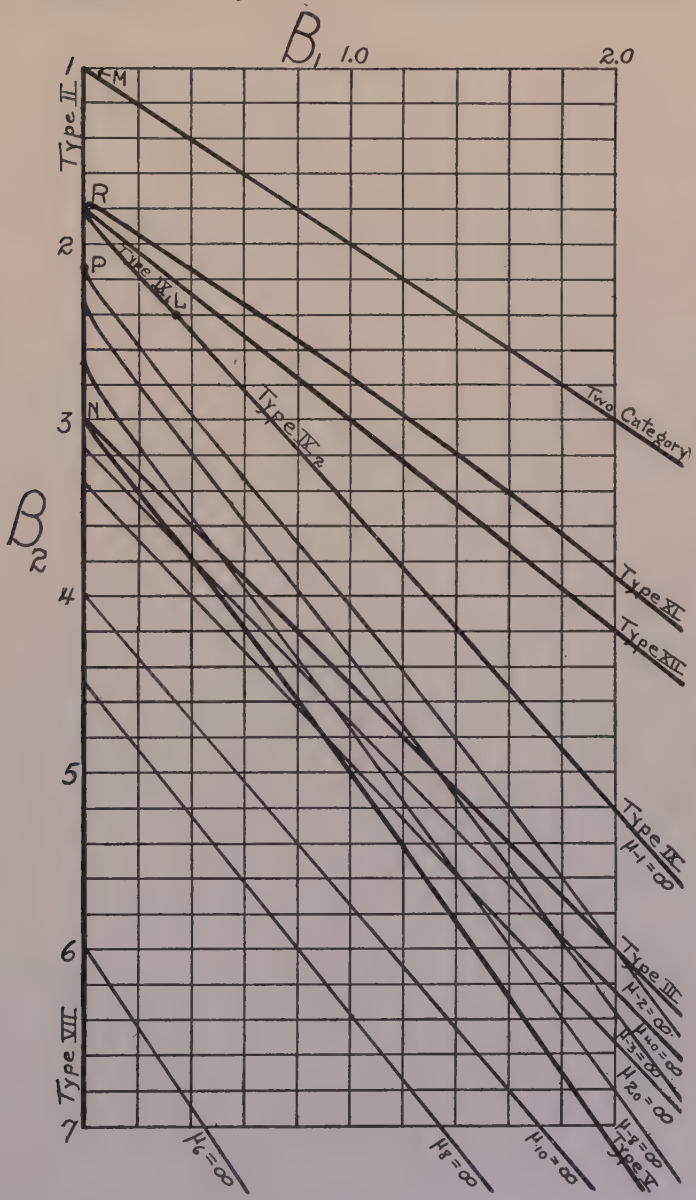
$$\kappa_2 = \frac{\beta_1\,(\beta_2 + 3)^2}{4\,(4\,\beta_2 - 3\,\beta_1)\,(2\,\beta_2 - 3\,\beta_1 - 6)} \qquad \text{(Criterion } \kappa_2) \ldots [72]$$

The connection between the β's and the type of curve may be shown by the illustrative curves of Chart XIX and by the following Chart XVIII which has in addition to the lines of Diagram XXXV in Pearson's Tables, certain lines and points for more recently discovered types of curves, as well as lines giving the finite limits of various moments. The meaning of the $(\mu_n = \infty)$ lines in Chart XVIII will be clear by an illustration. It is found by reference to the Chart that the lines $(\mu_{-8} = \infty)$ and $(\mu_{20} = \infty)$ approximately pass through the point $(\beta_1 = 1.45,\ \beta_2 = 5.66)$. The equation of the curve fitting a distribution yielding these β's has all of its moments between μ_{-8} and μ_{20} finite, and moments outside these limits are infinite. For the positive moments the mean, a finite boundary, or any other finite point, may be taken as the origin, while for the negative moments one of the boundaries of the distribution is the origin. For a point above Type III no positive moments are infinite and for a point below Type V no negative moments (defined further in Section 40) are infinite. Only certain of the breakdown lines, i.e., lines where the moment becomes infinite, have been drawn, there being an infinity of positive moment breakdown lines between $(\mu_{40} = \infty)$ and Type III and an infinity of negative moment breakdown lines between $(\mu_{-8} = \infty)$ and Type V. The discussion of the significance of these lines will follow shortly.

After determining β_1 and β_2 from the data, a corresponding point on Chart XVIII may be located. Should this be a point on a line the equation of the distribution will have two degrees of freedom in addition to that based upon N, the population. If the (β_1, β_2) point lies in a space between lines, the equation of the curve has one more constant in it and one greater degree of freedom. If the (β_1, β_2) point falls on certain designated spots on the lines, especially if it falls where two curves cross,

CHART XVIII

Diagram of Types of Frequency Distribution



the equation of the curve simplifies and has but one constant. In general the $(\beta_1\beta_2)$ point will not lie exactly on a line or on a unique point in a line, but if near such a place much labor in fitting a curve may be saved by choosing the simpler equation. This is frequently permissible, as may be decided from Charts and Tables given in Pearson's Tables, from which the probable error of the location of the $(\beta_1\beta_2)$ point may be determined. It is therefore possible to tell how unreasonable it would be to choose a type represented by the simpler form.

Section 38. Description of Types of Curves

We will first note points upon the lines which give the very simple one-constant equations. Reference to the drawings of Chart XIX will show the general form of the curves.

(*M*) The point of meeting of the line $\beta_1 = 0$, along which all distributions are symmetrical, and the line, $\beta_2 - \beta_1 - 1 = 0$, along which all distributions consist of frequencies in two categories.

$$\beta_1 = 0, \beta_2 = 1.0$$

At point (M) two equal categories constitute the distribution. Pearson has not given a name to this point nor assigned a type number to the line, $\beta_2 - \beta_1 - 1 = 0$. Due to the importance of the 1 : 1 ratio from the Mendelian point of view I have called this point (M). The line might be called the Mendelian line, but as it includes all two-category distributions and not simply those having Mendelian significance, I will call it the Two-Category Type Line.

(*R*) The point corresponding to a rectangular distribution.

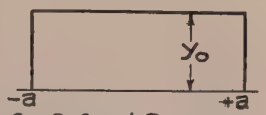
$$\beta_1 = 0, \beta_2 = 1.8$$

This point is the juncture of many lines and may therefore be considered a special case of any of the types which meet here, i.e., Types II-u, II-i, I-j, I-i, VIII, IX-1, XII. This point shares with the exponential the distinction of being the conflux of the greatest number of types of any point in the diagram, not excepting the normal point. There is a point, not in the field corresponding to real distributions $(\beta_1 = -4, \beta_2 = -3)$, which is still more exceptional as judged by the number of lines which pass through it.

M. Special Case of Type II u

Zero Base, or Zero Width Class Interval.

R. Rectangle.

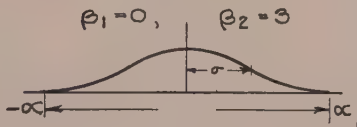


$\beta_1 = 0, \beta_2 = 1.8 \quad y = y_0$
Range from -a to +a

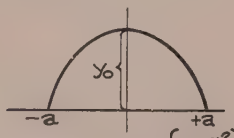
N. Normal or Gaussian.

$$y = \frac{N}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}$$

$\beta_1 = 0, \quad \beta_2 = 3$

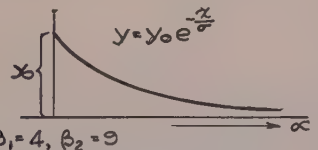


P. Parabolic.



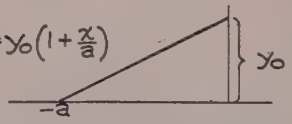
$\beta_1 = 0, \beta_2 = 2\frac{1}{7}, \quad y = y_0\left[1 - \frac{x^2}{a^2}\right]$
Range from -a to +a

E. Exponential: Type X



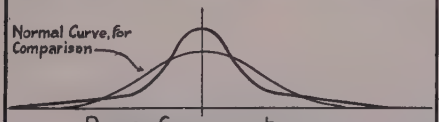
$\beta_1 = 4, \beta_2 = 9$
Range from 0 to ∞

L. Line: Point of Division Between Type IX 1 and IX 2



$\beta_1 = .32, \beta_2 = 2.4$ Range -a to 0

A. Curve drawn is $y = \frac{1}{\pi\sqrt{27}}\left(\frac{\beta_2 - 3}{3 + x^2}\right)^3$
Corresponding to Point $\beta_1 = 0, \beta_2 = 9$, and is Slightly Less Peaked Than Point A Curve



Range from -∞ to ∞

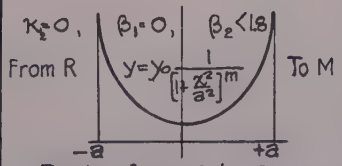
Type VII. See A.
$\kappa_2 = 0, \quad \beta_1 = 0, \quad \beta_2 > 3.0$

$$y = y_0 \frac{1}{\left(1 + \frac{x^2}{a^2}\right)^m}$$

Varies from N to a Curve More Leptokurtic than A
Range from -∞ to ∞

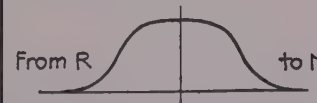
Type II-u

$\kappa_2 = 0, \quad \beta_1 = 0, \quad \beta_2 < 1.8$



Range from -a to +a

Type II - i



$\kappa_2 = 0, \beta_1 = 0, \quad 1.8 < \beta_2 < 3.0$
$y = y_0\left[1 + \frac{x^2}{a^2}\right]^m$

(*N*) The point corresponding to the normal distribution.

$$\beta_1 = 0,\ \beta_2 = 3.0$$

This is the conflux of Types I-i, II-i, III-i, IV-i, V, VI and VII. All of these are i-curves, that is, they are characterized by a single positive mode and have zero frequencies and a slope of zero at the upper and lower limits of the distribution. Further unique characteristics of this point will be pointed out in connection with reliability.

(*P*) The point corresponding to a parabola.

$$\beta_1 = 0,\ \beta_2 = 2\tfrac{1}{7}$$

This is simply a special point in the Type II-i line.

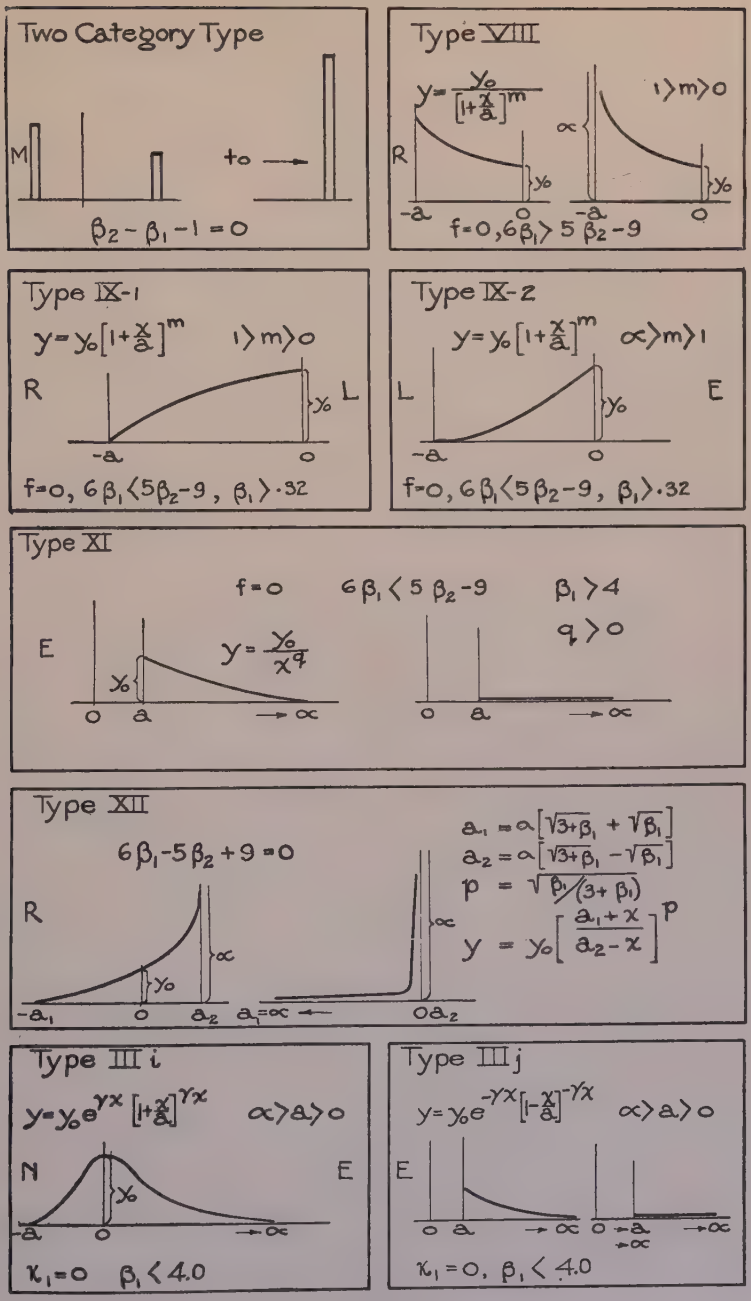
(*A*) The point corresponding to the symmetrical Type VII distribution for which the mean and the median are equally reliable averages. The point is not here located exactly, but it is in the neighborhood of

$$\beta_1 = 0,\ 11.0 < \beta_2 < 12.0$$

Below this point the median is more reliable than the mean and above this point less reliable. It should be noted that the line

$$8\,\beta_2 - 15\,\beta_1 - 36 = 0$$

is far above this point. The probable error of the fourth moment becomes infinity below this line. Accordingly the equation of a curve, or any other function involving the fourth moment, loses significance. The mean and the probable error of the mean do not involve a higher moment than the second, so that they remain significant for distributions for which it is impossible to fit a curve. In other words, the fourth moment breaks down as a significant feature of a distribution long before the second moment or the standard deviation; and these latter in turn break down before the first moment, or mean; and for certain distributions (e.g., $\beta_1 = 0$, $\beta_2 > 12.0$) the mean breaks down not only when the median does not, but when it is in fact rapidly improving as a measure of central tendency. Were we to go in the other direction into the Type II-u region we would find the median breaking down while the mean remains very reliable. This point is taken up later.


Two Category Type
M
to
$\beta_2 - \beta_1 - 1 = 0$
Type VIII
$y = \frac{y_0}{[1+\frac{x}{a}]^m}$
∞
$1 > m > 0$
R
y_0
-a
0
$f = 0, 6\beta_1 > 5\beta_2 - 9$
Type IX-1
$y = y_0[1+\frac{x}{a}]^m$
$1 > m > 0$
R
L
y_0
-a
0
$f = 0, 6\beta_1 < 5\beta_2 - 9, \beta_1 > .32$
Type IX-2
$y = y_0[1+\frac{x}{a}]^m$
$\infty > m > 1$
L
E
y_0
-a
0
$f = 0, 6\beta_1 < 5\beta_2 - 9, \beta_1 > .32$
Type XI
$f = 0$
$6\beta_1 < 5\beta_2 - 9$
$\beta_1 > 4$
$q > 0$
E
$y = \frac{y_0}{x^q}$
y_0
0
a
∞
Type XII
$6\beta_1 - 5\beta_2 + 9 = 0$
$a_1 = a[\sqrt{3+\beta_1} + \sqrt{\beta_1}]$
$a_2 = a[\sqrt{3+\beta_1} - \sqrt{\beta_1}]$
$p = \sqrt{\beta_1/(3+\beta_1)}$
$y = y_0\left[\frac{a_1+x}{a_2-x}\right]^p$
R
∞
y_0
$-a_1$
0
a_2
$a_1 = \infty$
Type III i
$y = y_0 e^{\gamma x}[1+\frac{x}{a}]^{\gamma x}$
$\infty > a_1 > 0$
N
E
y_0
-a
0
∞
$\gamma_1 = 0$
$\beta_1 < 4.0$
Type III j
$y = y_0 e^{-\gamma x}[1-\frac{x}{a}]^{-\gamma x}$
$\infty > a_1 > 0$
E
0
a
∞
$-\infty$
$\gamma_1 = 0, \beta_1 < 4.0$

(*L*) The point corresponding to the line distribution

$$\beta_1 = .32, \beta_2 = 2.4$$

This is a point of change of types. On the line to the left of this point distributions are Type IX-1 and to the right Type IX-2.

(*E*) The point corresponding to the exponential distribution

$$\beta_1 = 4.0, \beta_2 = 9.0$$

This point, which is well off the chart as drawn, is at the intersection of Type IX-2 and Type III lines. Type IX-2 curves become Type X curves at this point and Type XI curves beyond it. Type III-i curves become Type X curves at this point and Type III-j beyond it. The exponential is therefore located at the juncture of Types I-i, I-j, III-i, III-j, VI-i, VI-j, IX-2, XI.

There are at least five salient one-constant distributions, three of them, (*M*), (*R*) and (*N*), representing symmetrical distributions and two of them, (*L*) and (*E*), constituting division points on the one line that divides i from j curves.

Excepting the special points noted, points upon any of the lines in the diagram correspond to two-constant distributions.

Types II-u, II-i, VII. The line

$$\beta_1 = 0$$

represents three types, II-u, II-i, VII, in addition to the special points (*M*), (*R*), (*B*) and (*N*). Following Pearson, this line would be a boundary of "possible" distributions.

Two-Category Type. Another boundary would be the line

$$\beta_2 - \beta_1 - 1 = 0$$

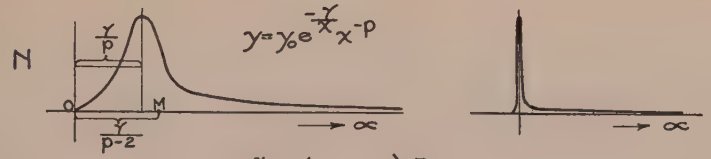
Looking upon distributions along this line as limiting cases of Type I-u distributions, it is seen that the equation representing them involves exponents which are infinite. For this reason no equation for this type is given.

Types VIII, IX-1, IX-2, XI. The line

$$\beta_1 (8\beta_2 - 9\beta_1 - 12)(\beta_2 + 3)^2 = (10\beta_2 - 12\beta_1 - 18)^2 (4\beta_2 - 3\beta_1)$$

represents Types VIII, IX-1, IX-2, XI in addition to the special points (*R*), (*L*) and Type X, or (*E*). This bi-quadratic, which we will call *f*, divides, on the one hand, the u-shaped

Type V



$\kappa_2 = 1 \qquad p > 5$

Type I-u
Boundaries: Two Category $\kappa_2 < 0$ Above $f=0$
Type II-u
Type VIII

$$y = y_0\left(1+\frac{x}{a_1}\right)^{m_1}\left(1-\frac{x}{a_2}\right)^{-m_2}$$

Type I-j
Boundaries: $f=0$ $\kappa_2 < 0$ Inside $f=0$

$$y = y_0\left(1+\frac{x}{a_1}\right)^{m_1}\left(1-\frac{x}{a_2}\right)^{m_2}$$

Type I-i
Boundaries: Types IX-1, IX-2, II-i, III-i $\kappa_2 < 0$ Below $f=0$

$$y = y_0\left(1+\frac{x}{a_1}\right)^{m_1}\left(1-\frac{x}{a_2}\right)^{m_2}$$

Type VI-i
Boundaries: Types III-i, XI, V $\infty > \kappa_2 > 1.00$ Below $f=0$

$$y = \frac{y_0 (x-a)^{m_1}}{x^{m_2}}$$

$m_1 > 0 \qquad m_2 > 0$

Type VI-j
Boundaries: Types III-j, XI $\infty > \kappa_2 > 1.00$ Inside $f=0$

$$y = y_0 \frac{(x-a)^{m_1}}{x^{m_2}}$$

$m_1 < 0 \qquad m_2 > 0$

Type IV
Boundaries: Types V, VII $1 > \kappa_2 > 0$

$$y = y_0 \frac{e^{-\nu \tan^{-1}\frac{x}{a}}}{\left(1+\frac{x^2}{a^2}\right)^m}$$

curves from the j-shaped, and on the other hand, the j-shaped from the i-shaped. All j-shaped curves lie within the arms of this bi-quadratic.

Type XII. The line

$$5\,\beta_2 - 6\,\beta_1 - 9 = 0$$

represents Type XII curves, which are j-shaped throughout the entire length of the line. In addition the special point (R) is on this line.

Types III-i, III-j. The line

$$2\,\beta_2 - 3\,\beta_1 - 6 = 0$$

represents Type III-i between points (N) and (E) and Type III-j beyond point (E). Containing as it does the two important points (N) and (E) and all points on the straight line connecting them, it is a very important type and, considering that it has but two parameters in addition to N, the population, it fits in a quite remarkable manner a large number of skew curves. Further characteristics of this type are pointed out later.

Type V. The line

$$4\,(4\,\beta_2 - 3\,\beta_1)\,(2\,\beta_2 - 3\,\beta_1 - 6) = \beta_1\,(\beta_2 + 3)^2$$

(Identical with $\kappa_2 = 0$)

represents Type V, composed entirely of i-shaped curves, in addition to the special point (N).

This completes the points and the lines. Points anywhere in the regions between lines correspond to three-constant distributions.

Type I-u. Composed entirely of u-shaped curves varying all the way from the Two-Category type to Type VIII.

Type I-j. Composed entirely of j-shaped curves. This region might appropriately be divided into two types, I-j-1 and I-j-2, depending upon which side of the Type XII line the point is located.

Type I-i. Composed entirely of i-shaped curves varying from Type IX to Type III. This is the only type area which is finite, as Type II, Type IX and Type III lines completely bound this region.

Types VI-i and VI-j. Type VI-i, composed entirely of i-shaped curves, lies below the Type III line and also below Type XI line. Type VI-j composed entirely of j-shaped curves, lies below Type III line and above Type XI line.

Type IV, composed entirely of highly leptokurtic i-shaped curves. This region lies below type V line. Below the line

$$8\,\beta_2 - 15\,\beta_1 - 36 = 0$$

is a region in which the probable error of the fourth moment is infinite, but it is not uncommon to find data which yield a (β_1, β_2) point below this line. In such case one of the outstanding features of the distribution is this very fact of an infinite eighth moment in the fitted curve, which is the cause of the infinite probable error of the fourth moment. Other significant features of the distribution may be determined from lower moments than the fourth, which continue to have finite probable errors for some distance below the critical line given. Pearson has named the region below this critical line the heterotypic region. As I understand the heterotypic to include bi-modal distributions I consider the designation inapt, as I can discover no evidence suggesting bi-modal tendencies in Type IV distributions. At present it is a sort of no-man's land. It is conceivable that there may be lines in it, corresponding to two-constant distributions not involving the fourth moment, and therefore determinable. There may also be unique points not involving either the third or the fourth moment. For one, the point $(\beta_1 = 0, \beta_2 = 9)$ may be considered such. The equation of this curve is

$$y = \frac{N}{\pi\sqrt{27}}\left(\frac{6}{3 + \frac{x^2}{\sigma^2}}\right)^3$$

It is the Type VII curve having the smallest possible integral exponent, and is completely determined by moments below the third and fourth. Furthermore, the probable error of the second moment, or standard deviation squared, is finite although the point $(\beta_2 = 9)$ is exactly twice as far down the Type VII line as the intercept $(\beta_2 = 4.5)$ of Pearson's critical line with the Type VII line. That this curve is not exceptional is obvious from the drawing of it given in Chart XIX, A.

Section 39. The Fitting of the Most Important Types of Curves

The normal distribution. The equation of this curve from the mean as origin is

$$y = \frac{N}{\sigma\sqrt{2\pi}} e^{\frac{-x^2}{2\sigma^2}}$$

The constants involved have been defined. The population, N, and the standard deviation of the distribution σ, are all that are needed to determine the normal curve which best represents given data.

Type II. The equation from the mean as origin is

$$y = y_0\left(1 - \frac{x^2}{a^2}\right)^m \quad \ldots\ldots\ldots\ldots\ldots\ldots [73]$$

in which

$$m = \frac{5\beta_2 - 9}{6 - 2\beta_2}$$

$$a = \tfrac{1}{2} \text{ the range} = \sqrt{\frac{2\mu_2\beta_2}{3 - \beta_2}}$$

$$y_0 = \text{ordinate at the mean} = \frac{N\,\Gamma(2m + 2)}{a\,2^{2m+1}\,[\Gamma(m + 1)]^2}$$

The Γ function may be evaluated without resorting to tables. First, if x is greater than 1, the following equation holds,

$$\Gamma(x + 1) = x\,\Gamma x \quad (\Gamma \text{ function reduction formula}) \ldots\ldots [74]$$

Second, if x is an integer greater than 1,

$$\Gamma(x + 1) = x! \quad (\Gamma \text{ function of an integer}) \ldots\ldots\ldots\ldots [75]$$

Third (Forsyth, quoted by Pearson 1901 supplement to 1895), as a close approximation to the value of the function, may be given,

$$\Gamma(x + 1) = \sqrt{2\pi}\left(\frac{\sqrt{\tfrac{1}{6} + x + x^2}}{e}\right)^{x+\frac{1}{2}} \quad \text{(Forsyth evaluation of the } \Gamma \text{ function)} . [76]$$

To quote from the reference cited, "If x be large the error is less than $1/(240\,x^3)$ of the whole." Even for an $x = 1.5$ the error is only in the neighborhood of 1 per cent. We may, however, first use the Γ reduction formula and then Forsyth's, for small values of x, resulting in as high a degree of accuracy as may be desired. For example,

$$\Gamma\,1.5 = \frac{\Gamma\,2.5}{1.5} = \frac{\Gamma\,3.5}{1.5 \times 2.5} = \frac{\Gamma\,4.5}{1.5 \times 2.5 \times 3.5} = \frac{\Gamma\,5.5}{1.5 \times 2.5 \times 3.5 \times 4.5}$$

The evaluation of $\Gamma\, 5.5$ by means of Forsyth's formula is highly reliable so that $\Gamma\, 1.5$ is readily obtained.

With the determination of y_0 the general solution of the Type II equation is completed.

Frequently, with immaterial loss in the excellence of fit, m may be set equal to the integer most nearly equal to $(5\,\beta_2 - 9)/(6 - 2\,\beta_2)$ and the resulting equation will be much simpler to plot. The use of an integral value for the exponent is equally serviceable in other types of curves. Whatever value of m is used as the exponent, is of course also to be used in the equation giving y_0.

Type VII. The equation from the mean as origin is,

$$y = \frac{y_0}{\left(1 + \frac{x^2}{a^2}\right)^m} \qquad [77]$$

$$m = \frac{5\,\beta_2 - 9}{2\,\beta_2 - 6} \qquad 2.5 < m < \infty$$

$$a = \frac{2\,\mu_2\,\beta_2}{\beta_2 - 3}$$

$$y_0 = \frac{N\,\Gamma\, m}{\sigma\sqrt{2\,\pi}\;\Gamma(m-\tfrac{1}{2})\sqrt{m-\tfrac{3}{2}}}$$

Note that μ_3 is not involved in the solution of the equations of Types II and VII. Types III and V do not involve μ_4.

Type III. The equation from the mode as origin is,

$$y = y_0\, e^{\frac{-px}{a}} \left(1 + \frac{x}{a}\right)^p \qquad [78]$$

$$\frac{p}{a} = \frac{2\,\mu_2}{\mu_3}$$

$$a = \mu_2\left(\frac{p}{a}\right) - \left(\frac{a}{p}\right)$$

$$y_0 = \frac{N\,p^{p+1}}{a\,e^p\;\Gamma(p+1)} = N\,\frac{p}{a}\,\frac{1}{\frac{\Gamma\,(p+1)}{e^{-p}\,p^p}}$$

$$\text{Mode} = \text{Mean} - \frac{a}{p}$$

Pearson (quoted in Duffell 1909) has shown that

$$\log\left(\frac{\Gamma\,(p+1)}{e^{-p}\,p^p}\right) = .3990899 + \tfrac{1}{2}\log p + .080929 \sin\frac{25°.623}{p} \qquad [79]$$

is a highly accurate equation for values of $p > 2$. It is accordingly a simple matter to determine y_0 by the aid of this equation.

A fitting of the distribution, not involving μ_3, may be accomplished by utilizing the fact that the difference between the mean and the mode equals a/p. Determine this distance by the use of formula [4] or [4-a], thus yielding a/p. The constants a, p, y_0 are then found as above, completing the solution.

Type V. The equation from the boundary as origin is,

$$y = y_0 e^{-\frac{\gamma}{x}} x^{-p} \qquad \text{.....[80]}$$

$$p = 4 + \frac{8}{\beta_1} + \frac{4\sqrt{\beta_1 + 4}}{\beta_1} \qquad \text{Plus sign of radical to be used}$$

$$\gamma = \sigma (p - 2) \sqrt{p - 3} \qquad \text{Sign of radical is the same as that of } \mu_3.$$

$$y_0 = \frac{N \gamma^{p-1}}{\Gamma (p - 1)}$$

$$\text{Distance from origin to mean} = \sigma \sqrt{p - 3}$$

$$\text{Mean} - \text{Mode} = \frac{2\gamma}{p (p - 2)}$$

Section 40. The Bearing of Curve Type upon Stability of Distribution

With the visual pictures of these curve types in mind we may proceed to a discussion of the bearing of type upon stability of distribution.

Mention has been made of the fact that the point ($\beta_1 = -4$, $\beta_2 = -3$) is a very unique point. The equation of every significant line in the chart except the line $\beta_1 = 0$, passes through this point. Many interesting relationships are made very clear by shifting the origin to this point.

The region enclosed within the Type II–VII and the Two-Category lines correspond to "real" distributions. A real distribution, as implied by the steps in the Pearson method, is one having the first four moments finite in addition to a finite total population. Other features, which one might insist should be finite, are not infrequently lacking. All of the u-shaped curves which are asymptotic to their upper or lower limits have infinite ordinates at these limits, though their areas are generally finite. One desirous of defining a real distribution in narrower terms than has Pearson would probably exclude these.

In speaking of infinite positive moments, ordinates or populations, the reader will of course understand that no obtained distribution can possess such a feature. Attempts to fit a smooth curve to a distribution more frequently than otherwise result in obtaining an equation with some infinite characteristic. Accordingly a reference to a distribution with such an infinite property is to the fitted curve, and though this infinite feature is not characteristic of the specific data in hand, it may be entirely descriptive of the total population of which the given data are a sample. In dealing with data in which certain reciprocal functions are infinite we will likewise be speaking of the fitted curves.

Certain of the Pearson types have infinite characteristics, ordinates, abscissas, and moments. As "real" distributions these might be looked upon as shortcomings. The point is, simply, that different limits as to the extent of distributions will exist dependent upon what is included in the concept "distribution." If negative frequencies are included, and it is to be hoped that a satisfactory physical meaning can be given to them so that they may be included,* then the limits of distributions greatly exceed the region bounded by the Type VII and the Two-Category lines. On the other hand, were one to restrict his concept to curves having finite eighth moments, the critical line ($\mu_8 = \infty$) would be a limit. The writer would think it logical either to restrict the concept to such as have all their moments finite, or to throw the field wide open and include everything which has as much as one determinable feature, such as the population, any one ordinate, any one moment, any one derivative, etc.

The acceptance of this broader definition of "distribution" immediately suggests the study of distributions for the purpose of ascertaining the nature and number of features which are finite, i.e. determinable. This has been done with reference to the moments of the various types of curves with results as shown in Chart XVIII. If a positive moment ($\int y\, x^n\, dx$) is finite when taken about a certain point, it continues finite when taken about any other point a finite distance from the

* For a suggestion as to this see Chart XIV and discussion of Sec. 8.

first. In dealing with negative, or inverse moments ($\int y\, x^{-n}\, dx$), however, the point of reference determines whether it be finite or infinite. The only natural point of reference seems to be a limit of the distribution. It is found, for all the Pearson curves, that if $\mu_{-n} = \infty$, then $\mu_{-(n+\Delta)}$, where Δ as well as n is positive, is of necessity also infinite, so that moments have been taken around that end of the distribution which shows a breakdown, or infinite value, in the lower inverse moment (μ_{-2} is called a lower, or smaller, negative or inverse moment than μ_{-3}, etc.).

The method of determining which are infinite follows from the fundamental differential equation, which is

$$\frac{dy}{y\,dx} = \frac{a_1 + a_2x}{c_1 + c_2x + c_3x^2}$$ (Pearson's differential equation for all types of uni-modal distributions)..[81]

If the roots of ($c_1 + c_2x + c_3x^2 = 0$) are imaginary the limits of the curve are $\pm\infty$, and if they are real the distribution lies between the values given by the roots. We may illustrate the method of determination of the moments by means of a Type I curve. To determine the infinite negative moment we will first shift the origin to the left extremity of the distribution.

$$\begin{aligned} \text{Let } c_3x^2 + c_2x + c_1 &= c_3\,(x + b_1)\,(x - b_2) \\ c &= b_1 + b_2 \\ a_2 &= b\,c_3 \\ a_1 - a_2b_1 &= a\,c_3 \\ z &= x + b_1 \end{aligned}$$

Then the equation from the new origin is

$$\frac{dy}{y\,dz} = \frac{a + bz}{-cz + z^2} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots[82]$$

and the limits are $z = 0$, and $z = c$. Multiplying by z^n and clearing give

$$\int ayz^n dz + \int byz^{n+1}\, dz = \int (-cz^{n+1} + z^{n+1} + z^{n+2})\, dy$$

Integrating,

$$aM_n + bM_{n+1} = [(-cz^{n+1} + z^{n+2})\, y]_{z=0}^{z=c} - \int -c\,(n+1)\, yz^n dz - \int (n+2)\, yz^{n+2} dz$$

$$[a - c\,(n+1)]\, M_n + (b + n + 2)\, M_{n+1} = [(-cz^{n+1} + z^{n+2})\, y]_0^c \ldots\ldots[83]$$

The M's or moments of this equation differ from the usual moments, μ's, only in that they are not divided by N, the

population. The two terms in the left hand member are functions of the entire distribution, while the right hand member is a function of the limits only. Whenever the coefficient $(b + n + 2)$ equals zero, then M_{n+1} can vary at will without affecting M_n. Therefore that value of n which makes this coefficient zero locates the moment, M_{n+1}, which becomes infinite. This is the procedure that could be followed in finding out where the positive moments break down, but in dealing with negative moments M_n becomes infinite before M_{n+1}, so that $[a - c(n + 1)]$ is then the coefficient that concerns us. It remains to express a and c in terms of β_1 and β_2.

Let $-a/c = m_1$ and $b = m_1 + m_2$, then the integral of the differential equation [82] is

$$y = kz^{m_1}(c - z)^{m_2} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots [84]$$

and the differential equation is,

$$-c(m_1 + n + 1)M_n + (m_1 + m_2 + n + 2)M_{n+1} = [(-cz^{n+1} + z^{n+2})y]_0^c \quad \ldots\ldots [83\,a]$$

If the origin is taken at the other boundary the differential equation is the same as above with m_1 and m_2 interchanged. The constants for any given distribution, m_1 and m_2, are functions of β_1 and β_2 (Pearson 1895) and can be expressed concisely if the following substitutions are made:

$$\gamma = \beta_1 + 4$$
$$\Delta = \beta_2 + 3$$
$$i = 4\Delta - 3\gamma$$
$$j = 5\Delta - 6\gamma$$
$$k = 3\gamma - 2\Delta$$

The two roots of the following equation give the two values of m,

$$km = j \pm \sqrt{\frac{9\beta_1\Delta^2(\Delta - \gamma)^2}{4ik + \beta_1\Delta^2}} \quad \ldots\ldots\ldots\ldots\ldots\ldots [85]$$

For the determination of the first inverse moment which breaks down we are concerned with the value found by using the minus sign of the radical. Values of m along a ray through the point $(\beta_1 = -4, \beta_2 = -3)$ may be readily determined. For example, for Type III line, $k = 0$, and

$$\beta_1 = \frac{4}{m + 1}$$

For Type XII line, $j = 0$, and

$$\beta_1 = \frac{3\,m^2}{1 - m^2}$$

For the line $11\,\gamma - 8\,\Delta = 0$,

$$\beta_1 = \frac{40\,(2\,m - 7)^2}{121\,(8 - m)\,(m + 1)}$$

As the M_n inverse moment breaks down when $n = -m - 1$, we may write for the Type III line, $\beta_1 = -4/n$, substitute -1, -2, -3, etc., values for n and ascertain the β_1's or the points along this line where the successive inverse moments become infinite. A similar procedure for other rays enables the plotting of the entire region, as shown in Chart XVIII.

Transferring the origin to the mean, so that positive moments will not become infinite merely due to the boundary being an infinite distance from the mean, and finding when the coefficient of M_{n+1} equals zero, gives the limiting values for the positive moments. These are more simple functions of β_1 and β_2, all being straight lines passing through the point ($\beta_1 = -4$, $\beta_2 = -3$). Going, on the chart, from below up, these rays become more and more dense until the limiting Type III ray is reached; just as, going from above down, the negative moment-breakdown lines become more and more dense until the limiting Type V line is reached. Special note needs to be made of the lines for moments μ_0, μ_1, μ_2, μ_3, and μ_4. The last three of these moments are incorporated in the very axes, β_1 and β_2, of the chart. Lines determined from the coefficient of M_{n+1}, showing where these moments break down, would show, as might have been anticipated, that the rays for μ_2, μ_3 and μ_4 lie outside of the region described by Pearson as that corresponding to real distributions. The line for ($\mu_0 = \infty$) when the coefficient of M_{n+1} is used lies within the Pearson possible region, and the line for ($\mu_1 = \infty$) lies at the boundary of it. The population, μ_0, is not necessary to the calculation of β_1 and β_2 so that the fact that it lies within this region is not inconsistent with the definition of the axes. However, μ_0 and μ_1 are smaller moments than those involved in β_1 and β_2 and it may be necessary to determine their points of breakdown from the coefficient of M_n and not of M_{n+1}.

Pending further study of Type I-u distributions I will not attempt an answer to this question or a description of distributions having infinite zero and first moments.

If the coefficient of M_{n+1} is examined with reference to the negative values of n for which it becomes zero, rays above Type III are located and these become more and more dense as Type III is approached. These have not been plotted, as earlier points of breakdown of the negative moments are located by dealing with the coefficient of M_n, but it is worth while noting that, *judging by the coefficient of* M_{n+1}, Type III distributions are the only ones which do not possess certain infinite positive or negative moments, i.e., certain elements of instability. If these unplotted lines should prove of any significance Type III distributions become unique not only because of possessing finite positive moments, but also because of the finite nature of whatever the inverse functions are whose points of breakdown are given by the coefficient of M_{n+1}. If, then, finite positive moments are of most importance III is the most stable of all the types; however, should finite negative moments be of greater importance than positive, Type V would be the most stable; and if the possession of both finite positive and negative moments is material then the normal distribution is the most stable curve within all the types.

It has for some time been known (Pearson, 1905), that if, by means of the first four moments, a curve is fitted to a distribution having a (β_1, β_2) point in region VI or IV, certain of the higher moments of the fitted curve are infinite. Pearson and Rhind (1909, pp. 130 and 134) have apparently interpreted this to mean that for such distributions moments higher than the fourth are needed for an adequate description of the data. This, however, hardly seems to me the most significant point of view. We can adequately and completely describe the sample collected by calculating and recording enough of the higher moments, but as Pearson has himself pointed out, this would scarcely yield valuable information as to the population of which the data are a sample because the probable errors of these higher moments become extreme. The really important conclusion to draw is that data, such that the

sample drawn gives a (β_1, β_2) point in the Type VI or IV regions, are of such a nature as to have indeterminate higher positive moments. The lines labeled $\mu_6 = \infty$, $\mu_8 = \infty$, $\cdots \mu_{-8} = \infty$, $\mu_{-3} = \infty$, etc., on Chart XVIII indicate where, judged by the first four moments, these higher positive and negative moments become infinite. Suppose that for a given (β_1, β_2) a fitted distribution is obtained for which $\mu_{20} = \infty$. Such analysis as I have been able to make leads me to infer that a few added moments in the fitting of the curve would not be expected to materially change this, and that some moment not far from μ_{20} will break down in any case.

These phenomena of instability of certain types of distributions are not mere oddities of the equations representing the types. Either coefficient of the difference equation connecting the moments may be written in the form,

$$\phi(\beta_1, \beta_2)\, n + f(\beta_1, \beta_2) = 0$$

in which ϕ and f are definite functions of the β's. Accordingly the breakdown of a moment is a function only of the moments involved in the β's. In other words, were we to fit a Type I curve and find that the n-th positive or negative moment became infinite, we could not improve the situation by fitting a Type II curve to the same data. The breakdown is not a function of the particular Pearson type chosen, but of the data, or of the differential equation back of all the Pearson types. That it is hardly the latter may be shown.

Had Pearson decided to use the first five moments in fitting curves it would have involved, in addition to the usual β_1 and β_2 constants, a third which we may call γ. A solid having three axes, β_1, β_2 and γ, would represent all the types just as the plane with axes β_1, β_2 now represents them all. The most serviceable function to constitute the third variable γ is not immediately obvious, but there would be certain advantages in defining γ as the difference between the β_3 $(\beta_3 = \mu_3\mu_5/\mu^4_2)$ given by the data and that derived from moments lower than the fifth by means of the present differential formula [81]. When so defined, if $\gamma = 0$ a distribution would be represented by a point on the two-dimensional (β_1, β_2) chart. It is barely conceivable that there might be a $(\beta_1, \beta_2, \gamma)$ line for which all

the positive and negative moments are finite. If there is such a line it cuts the (β_1, β_2) plane in the Normal point and nowhere else, so that the normal distribution loses none of its peculiar stability. The existence of such a line seems unlikely in view of the fact that there is no line (as opposed to point) in the (β_1, β_2) plane for which all the moments are finite. Otherwise expressed, had two moments only been used to derive the equations of curves, the special points on the chart could have been found and the normal distribution would have been the only one having all its moments finite. Had three moments been used the special lines in the chart could have been found, but no line would represent distributions having all their moments finite, the single Normal point again possessing this characteristic. Again, by the use of four moments, no area, no line, but merely the one Normal point is found for which all the positive and negative moments are finite. Accordingly it seems unlikely that the addition of a fifth moment would result in any extension of the distributions having all their moments finite.

The preceding discussion suggests that it would be futile to add an x_3 term in the denominator of the differential equation,

$$\frac{dy}{y\,dx} = \frac{a_1 + a_2 x}{c_1 + c_2 x + c_3 x^2}$$

The addition of an x^2 term in the numerator introduces bimodality and carries the problem into an entirely different field, corresponding, in all probability, to the operation of two opposing trends, instead of a single one such as we are here considering.

The only conclusion which seems to me to follow from the situation as described is that the weakness in distributions, evidenced by the existence of certain infinite moments in the fitted curves, lies in the data. This far reaching conclusion is supported by (1) the fact that an extension of the differential equation to include additional moments will, apparently, sometimes change, but not materially better the situation; and (2) by the known illustrations of instability which may be drawn from economic, psychologic and biologic fields.

Section 41. Illustrations of Unstable Distributions

Two distributions have come to my attention which are difficult to interpret, except as being unstable Type VI distributions.

The first is of price ratios, see Chart VIII, each ratio being the quotient of a price in a certain year divided by the price of the same commodity the preceding year. The distribution is very peaked and somewhat skewed and gives a (β_1, β_2) point so far down the chart that the fourth moment has an infinite probable error when the differential equation method of determining it is followed. The apparently puzzling question is how the curve fitting method can be so far wrong as to positively describe this distribution as one having an infinite feature. Recent study of similar price data shows that the fitted curve was undoubtedly correct and that the data did actually have such an infinite characteristic. Certain commodities for sale in 1917 were not purchasable at any price in 1918 and the series of 1918 ratios covered only such 1917 commodities as could be purchased in 1918. In other words, such price ratios as were recorded were in truth but a part of an unstable distribution, and being such they gave evidence that an occasional infinite price ratio was to be expected.

The second series is such as may be collected by any experimenter. A certain student was a subject in a reaction time experiment. The stimulus consisted of a spoken word and the reagent was directed to reply with the first word coming to mind. The series of reaction times revealed a Type VI distribution with a fourth moment having an infinite probable error when determined from the differential equation. This reagent was not tested further, but other reagents have been, with the result that a mental confusion or blocking has been found to occasionally occur, and to be so pronounced that the reagent has refused to react at all, i.e., the reaction time for that particular stimulus has become infinite. I have no doubt that were it possible, without changing the conditions, to continue the experiment with the first subject, sooner or later a similar blocking would be found, so that here again the probability is that the infinite higher moment is a true description of the situation.

According to Angell (1907), who points out that judgments of equality between two differing stimuli cease to constitute a homogeneous series if the stimuli differ by too great an amount, the same sort of condition holds generally in psychological threshold experiments. That is to say that reactions from such widely differing stimuli will yield distributions having unstable tails, or, what I would take as the statistical equivalent, Type IV or Type VI distributions. The use of the curve fitting method to determine the degree and nature of the instability in threshold experiments is suggested, but it suffices for our immediate purposes to note that psychologic as well as economic data occasionally yield distributions actually possessed of unstable tail functions, or in other words, infinite positive moments.

These illustrations point the possibility of the existence of a causal relationship which is determinable from a knowledge of the positive, and probably also negative, moments which become infinite. In fact, the order of the breakdown moments may prove a touchstone to the discovery of causal relationships. The method at present available for locating these critical moments is that of utilizing the first four positive moments from the actual data to determine a differential equation connecting moments. Having this equation the critical moments may be located immediately.

Slight shifting of the origin entirely changes the situation with reference to the inverse moments, so that, (*a*) it is either impossible to utilize inverse moments, (*b*) the conditions of the problem must give the limit with absolute definiteness, or (*c*) more definite features, such as the positive moments, must be used for the indirect determination of the limits and of the inverse moments around these limits. That method (*c*) will result in determinations with relatively small probable errors in case the lower negative moments are the critical ones is apparent from the appreciable distances apart of the μ_{-n} lines of Chart XVIII.

Though the laws controlling biologic phenomena have proven less easily and definitely determinable than many of those of physics, nevertheless the distributions of traits resulting from biological forces can readily be determined and examined. Is it not reasonable to think that, whatever else evolution may

involve it certainly involves a trend toward stability? If it is a development through laws represented by positive moments, its limit is a Type III distribution; and if through laws represented by inverse moments, its limit is a Type V distribution; and if both are involved, the only final limit is the normal distribution. This approach may be peculiarly valuable in studying evolution and it should not be a difficult matter to test it. Distributions of shell and skeletal structure of past ages can be made. Should it prove a fact that forms existent in the past giving distributions different from Types III and V have disappeared, and that those close to these Types are still represented by extant life, it would be complete support of this point.

We may note that the peculiar stability of Type III as judged by the existence of determinable positive moments is in harmony with the unique facts of correlation which Pearson has pointed out as belonging to this type. This is the only type in which "each contributory cause group is of equal valency and independent." The writer may have overlooked, but at least he has not found, in Pearson's contributions a satisfactory explanation and elaboration of "cause groups." He, however, interprets them as analogous to separate chromosomes, each of which may affect a single character, or to separate climatic and economic conditions each of which may affect a given food product, etc. If cause groups are not independent, so that a measure of a certain magnitude implies other magnitudes positively correlated with it, we have a situation which, from a priori considerations, one would expect to correspond to a trend, or tendency operating to pull measures in a certain direction, possibly entirely out of the distribution. It may be that a sufficient number of counteracting pulls, or vectors, could exactly balance each other, resulting in a condition identical with one not involving any pulls whatsoever, so that it seems equally reasonable to look upon Type III distributions as those in which there is a perfect balance between positive and negative correlation tendencies, thus revealing a zero correlation, or as distributions in which the pulls between elements are all zero. Whichever view is taken the significant result remains the same; that distributions which differ from

Type III thereby give evidence of the existence of uncompensated correlation between cause groups, — and of lack of stability since certain moments are indeterminate.

The determination of the specific nature of the correlation between cause groups in Type V distributions is a promising field of research. This type, holding as it does the same position with reference to stability of negative moments that Type III holds with reference to stability of positive moments, may possess some equally unique and stable characteristic with reference to negative product-moments as that possessed by Type III with reference to positive product-moments.

In the light of all the facts presented it would seem that evolution must be a trend toward the normal distribution. Also, dependent upon the causal forces operating, it would seem that subsidiary trends would be toward the three lines running into the normal point. If the causal forces can be expressed as positive moments, changes in distributions below Type III in the direction of Type III would mean ever greater stability, i.e., evolution. If the causal forces can be expressed as negative moments, changes in distributions above Type V in the direction of Type V would mean evolution. Balanced or symmetrical distributions show a peculiar stability in that all odd moments are zero. If stability of this type is the goal of a certain line of evolution, the trend would be toward Type II or Type VII. Finally, a certain development (biologic, economic, psychologic, or what not) having reached one of the three subsidiary goals, Type II or VII, Type III, or Type V, further advance, to insure stability of a still greater order, would be along the line toward the normal point.

The possession by an individual of a trait of such magnitude as to lie outside of the distribution given by the other members of the species ordinarily carries with it the elimination by death of the individual,* hence stability in trait is intimately connected with stability in species.

Only in case a trait is operated upon by such influences as result in the measures of the trait falling into a normal distribution can it be said that there is complete stability, or that the

* Cf. the traits possessed by *lethal drosophila melanogaster*.

race or species possessing it gives evidence of a self-contained permanence.

Clearly if this analysis is correct, the evolution of a bisexua' type of life would be as follows: (1) two entirely distinct traits which we may call male and female; (2) an occasional modification of the two, each in the direction of the other, giving a u-shaped distribution; (3) a building up of a common ground between the extremes, giving a limited range Type II-i distribution; and (4) a further weakening of the extreme characteristics until they become of infinitesimal importance in comparison with the common ground between, resulting in a normal distribution.

Following the lead of the argument we find the human species much further developed in certain parts of its makeup than in others. As illustrations of the four stages note (1) primary sex characteristics; (2) secondary sex characteristics; (3) musculature; (4) intelligence. In concluding this chapter let me emphasize the promise that lies in an experimental study of evolution, utilizing the facts of distribution types.

CHAPTER VIII

MEASURES OF RELATIONSHIP

Section 42. The Problem of Concomitant Variation in the Sciences

The determination of the law underlying concomitant variation is a problem common to all the sciences. The physical sciences have a great advantage over the social and biological sciences in that (1) errors of observation and measurement are usually very small in comparison with the measures involved and (2) fewer factors are ordinarily present. In measuring some intellectual capacity of a group of children, it usually happens that the probable errors of the test scores obtained are greater than half the standard deviation of the scores of the group. Obviously any relationship between two capacities, each measured with no greater reliability than this, will be clouded by the errors of measurement. This is serious enough, but it is not the only difficulty. In measuring the effect of gravity, physicists can ordinarily assume that ten pounds of lead and ten pounds of iron will act in a similar manner. But in measuring intellect, food prices, etc., to say that one reagent, one commodity, etc., is equivalent to another with respect to the function being examined, is usually questionable. Accordingly, where the investigations of physics lead to the establishment of "laws," those of the social sciences ordinarily lead to the discovery of "tendencies." Relationships between two psychological, biological or social factors frequently depend upon a number of causes, each more or less independent, and no one of which is so important as to dominate the situation. Under these conditions, the relationship tends to be rectilinear. In other cases, where the true relationship is not rectilinear, large errors of measurement will lessen the strength of the

measurable relationship, thereby making it more difficult to determine the exact nature of whatever curvilinear relationship may exist. It is also true that relationships which are intrinsically curvilinear when determined over a range of the two variables from very low to very high, may show practically rectilinear relationship throughout a short stretch of the range. For all the reasons stated, a measure of relationship based upon the assumption of rectilinearity is of great importance. Even in the case of known non-rectilinear relationship it is of much value as a point of departure. The balance of this chapter is devoted to a discussion of Pearson's product-moment coefficient of correlation, the "best" measure of mutual implication, if relationships are rectilinear.

The most fundamental properties of this measure of relationship were discovered and presented graphically by Francis Galton from 1877 to 1888. Galton's investigations had to do with the inheritance of traits, and certain of the terms which he used would hardly have arisen if the development had involved other data. For example, the symbol "r" was a measure of the "reversion," such, for example, as offspring upon mid-parent (a mid-parent measure is the average of the measures of father and mother). Later, Galton used the terms "regression" and "co-relation" and called the measure the "Index of Co-relation." Weldon very properly calls this measure "Galton's Function" and Edgeworth in 1892 gave it the name which has survived, "Coefficient of Correlation." Pearson (1920 notes) has pointed out that the product-moment function of Bravais bears but a resemblance in form to the product-moment coefficient of correlation. Whereas Bravais started with observations which were assumed to be independent, and in treating them obtained derived measures whose product-moments did not equal zero, Galton started with the epoch-making concept that the original measures were dependent. The Bravais treatment leads nowhere so far as correlation theory is concerned, because the measures which are correlated do not constitute original data, nor functions the correlations between which are of any moment on their own account. Partial correlation analysis leads to independent measures, having given related original scores;

which is exactly the reverse of the Bravais or Gaussian developments. Galton alone seems deserving of being called the father of correlation.

Section 43. Findings Resulting from Galton's Graphic Treatment

Galton s procedure, based upon medians and quartile deviations, has given way to the more accurate one involving the product-moment formula,

$$r = \frac{\Sigma xy}{N\sigma_1\sigma_2}$$

developed by Pearson.

We cannot do better than to use Galton's data in deriving a measure of correlation. Galton obtained the heights of parents and the heights of children, and drew up a "correlation table" or "scatter diagram" showing the relationship between the two. All female heights were multiplied by 1.08 to make them comparable with male heights. This procedure is not the most sound, but in this problem leads to no material error. Letting X_1, X_2 represent male and female heights, σ_1, σ_2, their standard deviations and M_1, M_2 their means, it would have been better to have reduced each female height to a comparable male height by the equation

$$\text{Comparable male height} = M_1 + (X_2 - M_2)\,\sigma_1/\sigma_2$$

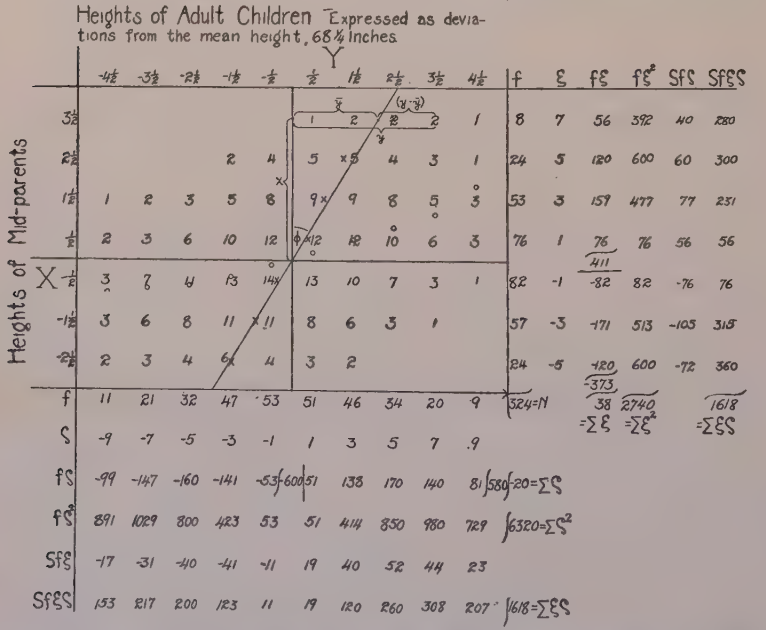
The discussion which follows will assume that the more reliable method of transmuting female into male heights was followed and also that the mean was used throughout. Presumably Galton used the median, but no fundamental difference in treatment followed from such use, it simply being a slightly less reliable procedure. Galton's diagram contained the data given in the accompanying correlation table or scatter diagram, Chart XX. Deviations being measured from $68\frac{1}{4}$ inches, which is a small fraction of an inch away from the true means, are labeled ξ and ζ instead of x and y, but no account of this slight difference is taken until the calculation of Section 45. From just such data as given, in fact it is likely that these identical data were involved, Galton inferred certain relationships which we now know hold with every normal correlation surface [Formula 87].

(*a*) A plot of the means of the vertical arrays (columns) as shown by the ×'s shows the "reversion" of offspring upon height of mid-parent. Thus if the mid-parent height is $2\frac{1}{2}$ inches above the mean the average or most probable height of offspring is $1\frac{1}{4}$ inches above the mean.

(*b*) The line connecting these means may be closely represented by a straight line through the origin or intersection of the means of the two distributions. This is the line showing the regression (or "reversion") of offspring upon mid-parent.

CHART XX

Correlation Between Heights of Mid-parent and Offspring



(*c*) There is a reversion or regression of mid-parent upon offspring. This would be represented by a straight line passing approximately through the o's. Thus for every correlation table there are two regression lines.

(*d*) The slopes of these two lines are equal, provided the standard deviations of the two distributions are equal.

(*e*) If standard deviations are equal, this slope varies between zero and one (Galton did not suggest the existence of negative correlations), and may be represented by the symbol "r".

(*f*) The standard deviations of the measures found in any one array (row or column) are approximately equal and are smaller than the standard deviations of the total distribution so that if σ_2 equals the standard deviation of the heights of offspring, and $\sigma_{2.1}$ the standard deviation of offspring corresponding to given heights of mid-parent, then

$$\sigma^2_{2.1} = \sigma^2_2 (1 - \lambda)$$

where λ is a positive quantity, also dealing with columns instead of rows,

$$\sigma^2_{1.2} = \sigma^2_1 (1 - \lambda)$$

in which λ is the same as before, σ_1 the standard deviation of heights of mid-parents, and $\sigma_{1.2}$ the standard deviation of heights of mid-parents corresponding to given heights of offspring. The symbol σ_a will, in subsequent formulas, stand for the standard deviation of an array around its own mean and $\sigma_{1.2}$ (or $\sigma_{2.1}$) the standard deviation of an array around the regression line, but as we are here dealing with homoscedastic rectilinear regression either symbol can be used, as $\sigma_a = \sigma_{1.2}$.

(*g*) There is a simple relation between λ and r.

It is, $\lambda = r^2$ so that

$$\sigma^2_{2.1} = \sigma^2_2 (1 - r^2)$$

and

$$\sigma^2_{1.2} = \sigma^2_1 (1 - r^2)$$

(Standard deviation of arrays from regression line, see also Section 48).....[86]

(*h*) Each array is approximately a normal distribution if the total distributions are normal.

(*i*) If contour lines for different frequencies are drawn in the diagram they constitute a system of similar and similarly placed ellipses, the conjugate diameters of which are the two regression lines.

Galton made no claim to mathematical ability but through sheer insight into the phenomena of mutual implication made these penetrating observations. He carried his conclusions, stated in probability terms, as to the nature of the correlation

surface, to Mr. J. D. Hamilton Dickson (1886), a mathematician, who readily wrote down the normal correlation equation involving two variables. In our present notation this is:

$$z = \frac{N}{2\pi\sigma_1\sigma_2\sqrt{1-r^2}}\, e^{\frac{-1}{2(1-r^2)}\left(\frac{x^2}{\sigma_1^2}+\frac{y^2}{\sigma_2^2}-\frac{2rxy}{\sigma_1\sigma_2}\right)}$$ (Normal correlation surface: 2 variables).... [see 88]

Galton's humility, after years of collection of data and subtle analysis of the same, in the face of the neat but not involved mathematical derivation, is worthy of note by the social scientists of this day who scoff at mathematical analysis. Upon receiving from Dr. Dickson the solution of his problem he wrote (quoted in Pearson 1920 notes), "I may be permitted to say that I never felt such a glow of loyalty and respect towards the sovereignty and magnificent sway of mathematical analysis as when his answer reached me, confirming, by purely mathematical reasoning, my various and laborious statistical conclusions with far more minuteness than I had dared to hope, for the original data ran somewhat roughly, and I had to smooth them with tender caution."

Section 44. Algebraic Statement of Galton's Graphic Findings and Derivation of Correlation Formulas

Let us consider these discoveries in more detail. Let x, the first variable, stand for height of mid-parent, y height of offspring, each expressed as a deviation from its respective mean. The standard deviations are respectively σ_1 and σ_2, while r is the slope of the regression line in the "reduced" scatter diagram, — that is, in the correlation table, — in which the measures entered are x/σ_1 and y/σ_2 respectively. Galton reduced by dividing by the quartiles, leading to essentially the same result as here. The slopes of the regression lines are equal, and equal to r. We will shortly obtain the numerical value of r by other than the graphic method of Galton. Finally, let $\bar{y}$ stand for an estimated height of offspring, knowing the height of mid-parent, and $\bar{x}$ the estimated height of mid-parent knowing height of offspring. With this notation, discoveries (*a*) and (*b*) together are equivalent to the equation

$$\frac{\bar{y}}{\sigma_2} = r\frac{x}{\sigma_1}$$ (Fundamental form of regression equation).....[see 91]

Propositions (*c*) and (*d*) are equivalent to the addition of the following equation to the preceding

$$\frac{\bar{x}}{\sigma_1} = r\frac{y}{\sigma_2} \quad \text{[see 91]}$$

These are the two fundamental regression equations characteristic of every regression table, showing rectilinear regression.

Proposition (*e*) is liable to misinterpretation. If $r = 0$, it implies that there is no relationship, no reversion or regression of one variable upon the other while an $r = 1$ means complete mutual implication of the two variables. More loosely stated, this latter situation will be described as one of complete mutual dependence, or simply dependence of the two variables. The student, however, should not postulate causal dependence. So far as data are concerned there is no evidence that the heights of the parents have any more to do in causing the heights of the offspring than do the heights of the offspring in causing the heights of the parents. This is characteristic of all measures of correlation. A situation exists and a correlation coefficient measures the tendency of the pairs of measures to be related but gives no evidence whether x is the cause of y, y the cause of x, or whether the cause is unknown and lies back of both. We think of parents being causal agents in determining the heights of offspring, but we do this for reasons outside of the scatter diagram, namely, the parents have existed earlier than the offspring in a time series.

Propositions (*f*) and (*g*) are of course the result of careful collection and study of data, but Galton gave a very simple proof of (*g*). The variability of the offspring generation is determined by the variability of the arrays (rows) and the variability of the means of these arrays. If Δ equals the distance of the mean of an array from the mean of the distribution and, as before, $\sigma_{2.1}$ equals the standard deviation of the array, and if n equals the number of measures found in an array, then $(n\sigma^2_{2.1} + n\Delta^2)$ equals the contribution of a single array in the calculation of the standard deviation, of the distribution, thus:

$$\sigma^2_2 = \frac{\Sigma\, n\sigma^2_{2.1} + \Sigma\, n\,\Delta^2}{N}$$

Since $\Sigma n\sigma^2_{2\cdot1} = N\sigma^2_{2\cdot1}$ and since for any array Δ equals the estimated y corresponding to the given value of x,

$$\Delta = \bar{y} = r\frac{\sigma_2}{\sigma_1}x$$

so that $\Sigma n\Delta^2 = \Sigma n\bar{y}^2 = N\sigma_{\bar{y}^2}$, therefore

$$\sigma^2_2 = \sigma^2_{2\cdot1} + \sigma^2_{\bar{y}}$$ (Standard deviation of distribution in terms of standard deviation of arrays and of standard deviation of means of arrays — rectilinear regression).....[87]

By proposition or discovery (f)

$$\frac{1}{N}\Sigma n\sigma^2_{2\cdot1} = \sigma^2_{2\cdot1} = \sigma_2(1-\lambda)$$

and

$$\frac{\Sigma n\Delta^2}{N} = r^2\frac{\sigma^2_2}{\sigma^2_1}\frac{\Sigma x^2}{N} = r^2\sigma^2_2$$

Accordingly,

$$\sigma^2_2 = \sigma^2_2(1-\lambda) + r^2\sigma^2_2$$

and finally,

$$\lambda = r^2$$

so that the important proposition (g) is established even before a formula for the arithmetical calculation of r is at hand.

(h) is an experimental finding which, coupled with (g) and (a), (b), (c) and (d), immediately gives the equation of the normal correlation surface. The equation, from the mean, of the normal distribution is,

$$z = \frac{1}{\sigma_2\sqrt{2\pi}}e^{\frac{-y^2}{2\sigma^2_2}}$$

If the distribution of an array is normal, its standard deviation $= \sigma_2\sqrt{1-r^2}$, and if its mean is $\Delta\ (= r\,\sigma_2/\sigma_1\,x)$ from the mean of the total population, then the equation of the normal distribution representing the array, from the mean of the entire distribution as origin, is,

$$z'' = \frac{1}{\sigma_2\sqrt{1-r^2}\sqrt{2\pi}}e^{\frac{-(y-r\sigma_2/\sigma_1 x)^2}{2\sigma^2_2(1-r^2)}}$$

The z'' corresponding to an assigned y is the probability of a measure *in* this array having the value y. The probability of a measure being in this particular x-array is

$$z' = \frac{1}{\sigma_1\sqrt{2\pi}}e^{\frac{-x^2}{2\sigma^2_1}},$$

Therefore, the probability of a measure having the particular value x and also the particular value y is the product of these two probabilities, $z = z'z''$,

$$z = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-r^2}} e^{\frac{-x^2}{2\sigma^2_1}} e^{\frac{-(y^2 - 2xyr\sigma_2/\sigma_1 + x^2r^2\sigma^2_2/\sigma^2_1)}{2\sigma^2_2(1-r^2)}}$$

which simplifies to,

$$z = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-r^2}} e^{\frac{1}{2(1-r^2)}\left(\frac{x^2}{\sigma^2_1} + \frac{y^2}{\sigma^2_2} - \frac{2rxy}{\sigma_1\sigma_2}\right)}$$

(Normal Correlation surface — 2 variables) . . [88]

This is the equation of the normal probability correlation surface of two variables and of a total population of one. If the right hand member is multiplied by N, we have the equation in case the total population is N. The quantity r has to this point been defined as the slope of the regression line in case standard measures [see formula 65] are the measures entered in the correlation table. We will now prove that in any scatter diagram, the two "best fit" rectilinear regression lines are

$$\frac{\bar{y}}{\sigma_2} = r\frac{x}{\sigma_1} \text{ and } \frac{\bar{x}}{\sigma_1} = r\frac{y}{\sigma_2} \quad \ldots\ldots\ldots\ldots \text{[see 91]}$$

in which the two r's are identical and given by the equation,

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2}\sqrt{\Sigma y^2}} = \frac{\Sigma xy}{N\sigma_1\sigma_2} \quad \ldots\ldots\ldots\ldots \text{[see 90]}$$

The term "best fit" is used as in the method of least squares. A "best fit" determination is one in which the sum of the squares or the errors of estimate is a minimum, that is, the standard error of estimate is a minimum. Determinations can be made resulting in the sum of the deviations; of the cubes of deviations; of their fourth powers, etc., being a minimum, but since the days of Gauss, it has been known that in the case of a normal distribution, none of these determinations will result in as small a median error as one in which the sum of the squares of the errors of estimate is made a minimum. The constants of distributions which are widely divergent from the normal, so determined that the standard error of estimate is a minimum, are undoubtedly very excellent determinations, but it is no longer possible to say that con-

stants so calculated have smaller median errors than would others derived upon a different principle. In all of the following treatment of simple and multiple correlation, the principle of least squares is involved and the standard errors are minimal, and because of this fact the determinations are called "best fit" determinations. They are "best" if the principle of least squares is the proper principle but they may not be so if some other principle is more sound, though in all cases we certainly can describe the least square as a highly excellent determination.

Referring to Chart XX, if the slope of the line drawn which is the regression of "y upon x," or the reversion of y toward x, is b_{21} (the numerical value of b_{21} is equal to tan ϕ) then having given a value x, the best estimate of the corresponding y value is $\bar{y}$. $\bar{y} = b_{21}x$. In general $\bar{y}$ will not be identical with the actual or experimentally obtained value of y, so that $(y - \bar{y})$ indicates an error of estimate. The standard error of estimate $\sigma_{2.1}$ is given by the equation,

$$\sigma^2_{2.1} = \frac{\Sigma\,(y - \bar{y})^2}{N}$$

The regression line which makes this magnitude a minimum is the regression line sought. Yule (1912) derives it without the use of calculus, but the calculus derivation is so much more simple that it is here given. See also in this connection problem 6 at end of this chapter.

$$f = \frac{\Sigma\,(y - b_{21}x)^2}{N} = \frac{\Sigma\,y^2 - 2\,b_{21}\Sigma\,xy + b^2_{21}\Sigma\,x^2}{N}$$

$$\frac{df}{db_{21}} = \frac{-2\,\Sigma\,xy}{N} + \frac{2\,b_{21}\Sigma\,x^2}{N} = 0$$

$$b_{21} = \frac{\Sigma\,xy}{\Sigma\,x^2} = \frac{\Sigma\,xy}{N\sigma_1{}^2}$$ (Regression coefficient of variable 2 upon variable 1, or the regression of the dependent variable, 2, upon the independent variable 1) .....[89]

This is the desired value of the regression coefficient. If standard measures are used the regression equation,

$$\bar{y} = \left(\frac{\Sigma\,xy}{N\sigma_1{}^2}\right) x$$

becomes

$$\frac{\bar{y}}{\sigma_2} = \left(\frac{\Sigma\,xy}{N\sigma_1\sigma_2}\right)\frac{x}{\sigma_1}$$

and the coefficient $\Sigma xy/(N\sigma_1\sigma_2)$ is the coefficient of correlation, r, or the measure of mutual implication, for a derivation similar to the preceding and involving the other regression line gives

$$b_{12} = \frac{\Sigma xy}{\Sigma y^2} = \frac{\Sigma xy}{N\sigma_2^2}$$ (Regression of variable 1 upon variable 2) .....[89]

so that

$$\frac{\bar{x}}{\sigma_1} = \left(\frac{\Sigma xy}{N\sigma_1\sigma_2}\right)\frac{y}{\sigma_2}$$

Thus the coefficient of correlation is given by

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2}\sqrt{\Sigma y^2}} = \frac{\Sigma xy}{N\sigma_1\sigma_2} = \sqrt{b_{12}b_{21}}$$ (Pearson product moment coefficient of correlation) ....[90]

and the regression equation may be written,

$$\frac{\bar{x}}{\sigma_1} = r\frac{y}{\sigma_2}$$ (Fundamental form of regression equation between two variables) .....[91]

The other regression is

$$\frac{\bar{y}}{\sigma_2} = r\frac{x}{\sigma_1}$$ .....[91]

Formula [91] may be written

$$\bar{x} = r\frac{\sigma_1}{\sigma_2}y, \text{ and } \bar{y} = r\frac{\sigma_2}{\sigma_1}x$$ .....[91 *a*]

or as

$$\bar{x} = b_{12}\,y, \text{ and } \bar{y} = b_{21}x$$ .....[91 *b*]

It is to be especially noted that whereas r_{12} always equals r_{21}, the regression coefficient b_{12} equals b_{21} only in case the two variables have equal standard deviations.

Section 45. The Detailed Steps in the Calculation of Correlation and Regression Constants

The steps necessary to the calculation of Σx^2, Σy^2 and Σxy are shown below and to the right of the diagram. (Chart XX.) The origin taken is $68\frac{1}{4}$ inches, but as shown by the sum of the fy row (– 20) and the sum of the fx column (38) the exact means are slightly different. We will calculate the correlation and regression coefficients without correcting for these slight discrepancies. They are taken into account in the calculation at the close of this section. To avoid working with fractions,

deviations from the means have been expressed in terms of one-half inch intervals. Thus a y-value of -9 means, 9 one-half inches below the mean. In terms of these units we have,

$$\Sigma \zeta^2 = 6320$$
$$\Sigma \xi^2 = 2740$$
$$\Sigma \zeta\xi = 1618$$

This last summation has been calculated in two ways so as to provide a check upon the arithmetical accuracy of the work. The first entry in the $\Sigma f\xi$ row is -17. This is the sum of the products of the frequencies of the ξ's for the single array for which $\xi = -9$. The notation, $Sf\xi$, is used to designate a summation for an array, whereas $\Sigma f\xi$, or more simply, $\Sigma\xi$, is the summation for the entire table. Similarly for $Sf\xi\zeta$ and $\Sigma\xi\zeta$. We have,

$$r = \frac{1618}{\sqrt{6320}\sqrt{2740}} = .3888$$

$$b_{21} = \frac{1618}{2740} = .5905 \quad \text{(Slope of regression line drawn)}$$

$$b_{12} = \frac{1618}{6320} = .2560 \quad \text{(Slope of other regression line)}$$

$$\sigma_1 = \sqrt{\frac{2740}{324}} = 2.908 \quad \text{(In one-half inch intervals)}$$
$$= 1.454 \quad \text{(In inch intervals)}$$

$$\sigma_2 = \sqrt{\frac{6320}{324}} = 4.417 \quad \text{(In one-half inch intervals)}$$
$$= 2.203 \quad \text{(In inch intervals)}$$

The regression of height of offspring upon mid-parent, in inch units, and measured from an origin of $68\frac{1}{4}$ inches, is,

$$\frac{\bar{\zeta}}{2.208} = .3888\,\frac{\xi}{1.454}$$

or,

$$\bar{\zeta} = .5905\,\xi$$

Having the equation in this fundamental form it is but a step to express it in terms of gross scores. Letting M' = the arbitrary origin or approximate mean, we have

$$\bar{\zeta} = \bar{Y} - M'_2$$
$$\bar{\xi} = \bar{X} - M'_1$$

Accordingly,

$$\bar{\zeta} = b_{21}\xi$$

may be written

$$\overline{Y} - M'_2 = b_{21}(X - M'_1)$$

or

$$\overline{Y} = b_{21}X + (M'_2 - b_{21}M'_1)$$
$$\overline{X} = b_{12}Y + (M'_1 - b_{12}M'_2)$$

To illustrate the use of this equation let us estimate the most probable height of male offspring if the mid-parent height is 64 inches.

$$\overline{Y} = .5905X + [68.25 - (.5905)(68.25)]$$
$$= .5905X + 27.95$$

Solving, when $X = 64$, gives, $Y = 65.74$, as the most probable height of offspring, or the mean height of many such offspring.

The calculation of the constants involved in the regression equation as shown assumes that deviations are from the means of the two distributions. In case origins other than means are used corrections may be applied to secure the product moment and standard deviations from the means. The corrections for the standard deviations have already been given, formula [22].

Let $\Delta_1 = M'_1 - M_1$, the distance from the arbitrary origin to the mean of the X's, and let $\Delta_2 = M'_2 - M_2$. Then,

$$\sigma^2_1 = \frac{\Sigma x^2}{N} = \frac{\Sigma \xi^2}{N} - \Delta^2_1$$

$$\sigma^2_1 = \frac{\Sigma y^2}{N} = \frac{\Sigma \zeta^2}{N} - \Delta^2_2$$

$$\Sigma xy = \Sigma(\xi + \Delta_1)(\zeta + \Delta_2)$$
$$= \Sigma(\xi + \Delta_1)\zeta + \Delta_2\Sigma(\xi + \Delta_1)$$
$$= \Sigma \xi\zeta + \Delta_1\Sigma\zeta + \Delta_2\Sigma(\xi + \Delta_1)$$

and since $\Sigma(\xi + \Delta_1) = \Sigma x = 0$ and $\Sigma\zeta = -N\Delta_2$, therefore,

$$\Sigma xy = S\xi\zeta - N\Delta_1\Delta_2$$ (Formula for correction of product — moment due to use of arbitrary origins) .....[92]

Accordingly r may be calculated from any origins whatever by the formula

$$r = \frac{\Sigma \xi\zeta - N\Delta_1\Delta_2}{\sqrt{\Sigma \xi^2 - N\Delta^2_1}\sqrt{\Sigma \zeta^2 - N\Delta^2_2}}$$ (Pearson product — moment coefficient of correlation calculated from any origin) .[93]

When zero is, for each variable, the arbitrary origin the above formula is equivalent to

$$r = \frac{\Sigma XY - (\Sigma X)(\Sigma Y)/N}{\sqrt{\Sigma X^2 - (\Sigma X)^2/N}\sqrt{\Sigma Y^2 - (\Sigma Y)^2/N}}$$ (r calculated from zero as arbitrary origin) .....[94]

Another variation is

$$r = \frac{\Sigma XY - NM_1M_2}{\sqrt{\Sigma X^2 - NM^2_1}\sqrt{\Sigma Y^2 - NM^2_2}}$$ (r calculated from zero as arbitrary origin).....[95]

Similarly,

$$b_{12} = \frac{\Sigma XY - NM_1M_2}{\Sigma Y^2 - NM^2_2}; \quad b_{21} = \frac{\Sigma XY - NM_1M_2}{\Sigma X^2 - NM^2_1}$$

(Regression coefficients calculated from zero as arbitrary origin).....[96]

Thus, for Galton's data, the correct values for the requisite constants are

$$r = \frac{1618 - (38)(-20)/324}{\sqrt{2740 - (38)^2/324}\sqrt{6320 - (-20)^2/324}} = .3897$$

$$b_{21} = .5923$$
$$M_1 = 68.25 + 38/324 = 68.37$$
$$M_2 = 68.25 - 20/324 = 68.19$$
$$\sigma_1 = 2.906$$
$$\sigma_2 = 4.416$$

Thus the corrected regression equation from the actual means as origins is

$$\bar{y} = .5923\, x$$

which differs but slightly from that obtained neglecting Δ_1 and Δ_2,

$$\bar{\zeta} = .5905\, \xi$$

and the corrected regression equation from zero inches as origins is

$$\bar{Y} = .5923\, X + 27.69$$

which in turn differs but slightly from that obtained neglecting Δ_1 and Δ_2.

Section 46. The Error Involved in Certain Approximations

It is desirable to know how large an error in the means may safely be neglected. We have, letting $s_1 = (\Sigma\xi^2)/N$ and $s_2 = (\Sigma\zeta^2)/N$,

$$r = \frac{\Sigma\, \xi\zeta - N\Delta_1\Delta_2}{N\sqrt{s_1{}^2 - \Delta_1{}^2}\sqrt{s_2{}^2 - \Delta_2{}^2}}$$

and we wish to ascertain how greatly this differs from the approximate value,

$$r' = \frac{\Sigma \xi\zeta}{Ns_1s_2}$$

Setting the expressions $1/\sqrt{s^2 - \Delta^2}$ equal to $1/s\sqrt{1 - \left(\frac{\Delta}{s}\right)^2}$ and expanding the radical by the binomial theorem, discarding powers in Δ/s greater than the second as being negligible in comparison with the second powers, gives, after certain simple reductions,

$$r = r' + r'\left[\frac{1}{2}\left(\frac{\Delta_1}{s_1}\right)^2 + \frac{1}{2}\left(\frac{\Delta_2}{s_2}\right)^2\right] - \frac{\Delta_1\Delta_2}{s_1s_2} \quad \text{(Showing error in } r \text{ from use of approximate means) [97]}$$
$$= r' + \epsilon$$

in which ϵ is the error introduced in case r' is taken as the value of r. Note that if r' is positive, less error is introduced if Δ_1 and Δ_2 have the same sign than if they are of opposite sign. Let us assume the two magnitudes (Δ/s) are equal. Then,

$$\epsilon = (r' - 1)\left(\frac{\Delta}{s}\right)^2$$

In this case ϵ is negative, i.e., if approximate means which are in error in the same sense are used, the obtained correlation, r', is larger than the correct value, r. We may solve the preceding equation for Δ/s for assigned values of r' and ϵ. The following tables give certain solutions:

If Errors in Means are Equal and of the Same Sign				If Errors in Means are Equal and of Opposite Sign			
ϵ	r'	Δ/s	Δ = approximately	ϵ	r'	Δ/s	Δ = approximately
−.001	.0	.032	1/158 of range	.001	.0	.032	1/158 of range
−.005	.0	.071	1/ 71 " "	.005	.0	.071	1/ 71 " "
−.010	.0	.100	1/ 50 " "	.010	.0	.100	1/ 50 " "
−.001	.7	.058	1/ 87 " "	.001	.7	.024	1/207 " "
−.005	.7	.129	1/ 42 " "	.005	.7	.054	1/ 92 " "
−.010	.7	.183	1/ 27 " "	.010	.7	.077	1/ 65 " "
−.001	.9	.100	1/ 50 " "	.001	.9	.023	1/218 " "
−.005	.9	.224	1/ 22 " "	.005	.9	.051	1/103 " "
−.010	.9	.316	1/ 16 " "	.010	.9	.073	1/ 69 " "

Since for Δ's of a given size, there is much greater error in the correlation coefficient if they are of opposite sign than if of

the same sign, therefore in choosing arbitrary means, it is frequently desirable to so choose that Δ_1, Δ_2 have the same sign. For example: suppose $\sigma_1 = \sigma_2 = 3$, $M_1 = 12.56$ and $M_2 = 9.30$, then better results will be obtained, if correction for arbitrary means is not made, by choosing 12.0 and 9.0 ($\Delta_1 = .56$ and $\Delta_2 = .30$) than by choosing 13.0 and 9.0 ($\Delta_1 = -.44$ and $\Delta_2 = .30$). For many investigations, an error of 1 per cent is not material so that, as a practical procedure subject to refinement if low correlations are involved or if a 1 per cent error is serious, it is safe to forego correcting for arbitrary means if the error in each of the means is less than 1/27 of the range and if they are of the same sign. This requirement is more easily met than one imposing the condition that the standard deviation should not be in error by more than 1 per cent. As standard deviations are usually features of a distribution which it is desirable to know, it seems better to forego correction for an arbitrary mean only in case the error introduced in the standard deviation is less than 1 per cent. We have

$$\sigma = \sqrt{s^2 - \Delta^2} = s - \tfrac{1}{2}\frac{\Delta^2}{s} + \text{higher powers in } \left(\frac{\Delta}{s}\right).$$

The error introduced by using s in place of σ is

$$s - \sigma = \frac{\Delta^2}{2\,s} \quad \text{(Showing error in } \sigma \text{ from use of approximate mean)} \ldots\ldots [98]$$

and the proportionate error is

$$\frac{s - \sigma}{s} = \frac{\Delta^2}{2\,s^2}$$

If an error of 1 per cent is permissible, we may write

$$\frac{\Delta^2}{2\,s^2} = .01$$

$$\frac{\Delta}{s} = .1414$$

or Δ is approximately 1/35 of the range. If there are 18 or more intervals in the range covered by the measures and if the arbitrary mean is chosen as the middle of the interval nearest the correct mean, then the error will be less than $\frac{1}{2}$ the interval or less than 1/36 of the range, so that the error in the resulting standard deviation will be less than 1 per cent

The correction just considered is on account of displacement

of the mean. Sheppard's correction, formula [68 *a*], is for grouping. If σ equals the correct standard deviation and S the standard deviation obtained from coarsely grouped data, Sheppard's correction gives

$$\sigma^2 = S^2 - 1/12$$

$$\sigma = S - 1/24\,S + \text{higher powers in } (1/12\,S)$$

$$\frac{S-\sigma}{S} = 1/24\,S^2 \quad \text{(Showing error in } \sigma \text{ due to grouping)} \ldots [99]$$

and if this equals .01, we have

$$S = 2.041$$

If the standard deviation is 2 or a trifle greater, the range is in the neighborhood of 10 or 12, so that if we have as many as 12 steps, the error of the standard deviation due to grouping is less than 1 per cent. The most exacting condition is therefore the one preceding this.

Accordingly, if there are 12 or more intervals in the ranges of both variables, and if the origins are so taken, by resorting to $\frac{1}{2}$ or $\frac{1}{3}$ steps if necessary, as not to differ from the correct means by more than $1/25$ of the range if the correlation is above .70, or $1/50$ if near .00; and if the origins taken lie either both above or below the correct means; the error introduced in either the standard deviations or the coefficient of correlation by not correcting for grouping or for approximate means, is less than 1 per cent. In case intervals are of necessity so broad that a material error in correlation results, the raw correlation coefficient requires a correction for broad categories.

Section 47. The Bearing of Broad Categories upon Correlation

Writing p_{11} for the product moment $\Sigma xy/N$, as in Section 48, we have

$$r_{12} = \frac{p_{11}}{\sigma_1 \sigma_2}$$

Ordinarily σ_1 and σ_2 will be taken as the standard deviations of the class indexes, but more accurate values are obtained by first applying Sheppard's corrections, formula [68 *a*]. Thus if h and k are the group intervals, s_1 and s_2 the standard deviations before applying Sheppard's corrections, we have,

$$\sigma^2{}_1 = s^2{}_1 - \frac{h^2}{12}, \text{ and } \sigma^2{}_2 = s^2{}_2 - \frac{k^2}{12} \ldots\ldots\ldots\ldots [68\,a]$$

To a first approximation there is no correction for grouping to be made to the product moment, p_{11}, so that we have

$$r_{12} = \frac{p_{11}}{\sqrt{s^2_1 - \frac{h^2}{12}}\sqrt{s^2_2 - \frac{k^2}{12}}} = \frac{\Sigma xy}{N\sqrt{s^2_1 - \frac{h^2}{12}}\sqrt{s^2_2 - \frac{k^2}{12}}}$$

(Coefficient of correlation after applying Sheppard's corrections) . . . . . [100]

If the grouping is very coarse and irregular we may assume a normal distribution and determine the *mean* of each class; calculate the correlation, using these mean class values as our variates, and correct for grouping. The correction for grouping is different from Sheppard's because here our correction is on account of using mean-class-values in place of the continuous variate, whereas Sheppard's correction is on account of using mid-class-values in place of the continuous variate. To point the distinction the following hypothetical problem involving trade ratings and general intelligence ratings is given.

	Ratings of General Mental Ability				Per Cent in Each Class	q, — Proportion Above Class	z, — Ordinate at Upper Limit of Class	x, — Mean of Class
	Dull	*Average*	*Bright*			.00	.000000	
Expert	1	4	5	10	10			1.755
						.10	.175498	
Journeyman	4	10	16	30	30			.703
						.40	.386342	
Apprentice	4	11	5	20	20			.000
						.60	.386342	
Novice	11	25	4	40	40			−.966
	20	50	30	100	.	1.00	.000000	
z'	.000000	.279962	.347693	.000000				
y		−1.400	−.135	1.159				

$$\sigma_x = \sqrt{80.40968/100} = .896714$$

$$\sigma_y = \sqrt{82.95276/100} = .910784$$

$$r_{yx} = \frac{28.57438}{100 \times 89671 \times .910784} = .34987$$

The symbols q, z and x have the meanings of Section 27. Formula [55] is used in determining x from q and z. Treating x and y as the values of the deviates from the means, the correlation coefficient is, by the usual process, found to equal .34987. This value, however, suffers from a large grouping error. We cannot apply Sheppard's corrections because we do not have equal class intervals and because we have not dealt with class indexes, but class means. Whereas s, the standard deviation of class indexes, is greater than σ, the standard deviation of the continuous variates; s', the standard deviation of class means, is less than σ. If class intervals are equal and equal to the unit used, we have,

$$\sigma^2 = s^2 - \frac{1}{12} \qquad \text{(Sheppard's correction)} \ldots [68\ a]$$

also

$$\sigma^2 = s'^2 + \frac{1}{12} \qquad \text{(Pearson's correction to the standard deviation of class means)} \ldots [101]$$

This second formula, as well as subsequent ones in this section, was derived by Pearson (1913 meas.). We thus see that an entirely different correction is needed. This last correction is not of general utility, as the problems in which we use class means instead of class indexes are usually such that we do not know that the class intervals are equal. We may, however, determine the correction by aid of the correlation between the variate and the class means of the classes into which the variates are placed.

Let χ be the value of the continuous variate, and x the value of the means of the classes into which the χ measures are placed. Then the regression of the χ's upon the x's is

$$\bar{\chi} = r_{\chi x} \frac{\sigma_\chi}{\sigma_x} x$$

but $\bar{\chi}$ is the mean of all the variates in the class of which x is the class mean, or simply $\bar{\chi} = x$. Substituting in the preceding equation we have

$$r_{\chi x} = \frac{\sigma_x}{\sigma_\chi} \qquad \text{(Correlation between a variate and the means of the classes in which it is recorded)} \ldots [102]$$

The standard deviation of the class means σ_x is obtained by calculation, and σ_χ is known if the form of distribution is

known. For the problem given $\sigma_x = .896714$ and $\sigma_\chi = 1.0$ since a normal distribution of standard deviation equal to 1.0 was assumed, so that $r_{\chi x} = .8967/1.0 = .8967$, and if γ stands for the continuous variate in the case of the second variable, then $r_{\gamma y} = .9108/1.0 = .9108$.

Continuing we may find the correlation between two continuous variates when each is recorded in broad categories. The following simple derivation depends upon principles of partial correlation discussed in Chapter XI. The reader should therefore be familiar with that chapter before attempting to follow this proof. The symbol $r_{xy.\chi\gamma}$ stands for the correlation between class means for constant values of the graduated variates χ and γ. Clearly when χ and γ are constant the corresponding class means x and y do not vary, so that $r_{xy.\chi\gamma} = 0$. This partial correlation coefficient, $r_{xy.\chi\gamma}$ is equal to a numerator determinant divided by the square root of the product of two others. The divisor is easily shown to be intrinsically positive so that the quotient becomes zero with the dividend. Accordingly we have

$$\begin{vmatrix} r_{xy} & r_{\chi y} & r_{\gamma y} \\ r_{\chi x} & 1 & r_{\chi\gamma} \\ r_{x\gamma} & r_{\chi\gamma} & 1 \end{vmatrix} = 0$$

in which $r_{\chi\gamma}$ is the corrected value sought, and r_{xy} is the value calculated, using the means of the broad categories. It has just been shown that $r_{\gamma y}$ is equal to σ_y/σ_γ, and $r_{\chi x}$ equal to σ_x/σ_χ. We need to know $r_{\chi y}$ and $r_{x\gamma}$. The partial correlation $r_{\chi y.\gamma}$ is that between the variate χ for a given value of the second variate γ, and the class mean y for a given value of the second variate γ. The class mean for a given value of the variate is invariable, so that y for constant γ is constant and accordingly $r_{\chi y.\gamma} = 0$. This partial coefficient can be zero only when the numerator of the quotient which is equal to it is zero; that is

$$r_{\chi y} - r_{\chi\gamma} r_{\gamma y} = 0$$

or

$$r_{\chi y} = r_{\chi\gamma} r_{\gamma y}$$

Similarly

$$r_{x\gamma} = r_{\chi\gamma} r_{\chi x}$$

Substituting in the determinant and solving for $r_{X\Upsilon}$ gives, if we let $_m r_{xy} = r_{X\Upsilon}$

$$_m r_{xy} = \frac{r_{xy}}{r_{Xx} r_{\Upsilon y}} = \frac{r_{xy}\sigma_X\sigma_\Upsilon}{\sigma_x\sigma_y}$$ (Giving the correction to r on account of use of class means). . . [103]

In case a normal distribution of standard deviation 1 is assumed to fit the distributions of the two variables, and the means of categories calculated upon this assumption, σ_X and σ_Υ each equal 1 so that we have

$$_m r_{xy} = \frac{r_{xy}}{\sigma_x\sigma_y} = \frac{\Sigma\, xy}{N\sigma^2{}_1\sigma^2{}_2}$$ (Correction to r on account of use of class means, upon assumption of a unit normal distribution). . [104]

The correction here derived for broad categories is equally serviceable when determining correlation ratios or contingency coefficients as described in Section 68.

Note that there are two corrections; one, Sheppard's, to be applied on account of broad equal intervals when class indexes are taken as the variates; and the second, the one here given, to be applied when the class means of broad equal or unequal intervals are taken as the variates. No correction is as yet worked out for application when class indexes are used and the intervals are broad and unequal, though in such case good results may be expected by empirically setting h in Sheppard's formula [68 a] equal to the mean of the several intervals involved.

We may return to the numerical problem and apply the correction to obtain the correlation corrected for broad categories between trade ratings and estimates of intelligence. It yields

$$\frac{.34987}{.896714 \times .910784} = .4284$$

In this calculation it has been assumed that x is the same for the first cell (expert-dull), the second cell (expert-average), and the third cell (expert-bright), and similarly throughout the rest of the table. This is only approximately true, and in case the categories are very broad and the correlation high it is far from true. The method should not be used with a four-fold table and it is of doubtful validity for the table given. It may be applied with good results if no class contains more than 25 or 30 per cent of the cases and if the correlation is not greater than .9.

Section 48. Properties of Correlation Surfaces

With the scatter diagram of Chart XX before us, the meanings of certain terms will be readily grasped. If the standard deviation of the successive x arrays are equal, the distribution is homoscedastic in the x variable and if, in addition, the standard deviations of the y arrays are equal, the correlation surface is homoscedastic in both senses. If the slope of the distribution in an array a given distance above the mean of the array, is equal to the slope the same distance below, and if this is true of all arrays, the total distribution is called homoclitic; thus, a distribution composed of symmetrical arrays is homoclitic. If means of successive arrays lie in a straight line, the regression is rectilinear or, by some writers, is termed linear. In case a regression table is homoscedastic, homoclitic, and has two rectilinear regression lines, the most probable value of one variable when estimated from a knowledge of the other, is that given by the regression equation. The regression determination in the case of distributions showing moderate divergence from these three conditions will still be very nearly the most probable. Scatter diagrams showing extreme divergence should be treated by some other method. Lack of substantial rectilinearity in regression is the most readily detected feature of a correlation surface which vitiates the use of the product moment coefficient of correlation. For most problems, the establishment of rectilinearity is sufficient to completely justify the use of the Pearson product moment coefficient of correlation. Note that this is a much easier requirement to meet than that the correlation surface be normal, that is, capable of accurate representation by means of equation [89]. Accurate correlation results may regularly be expected from distributions showing rectilinear regression lines, but otherwise widely divergent from the normal correlation surface. Due to the fact that Pearson's early development of the product moment coefficient of correlation was based upon the assumption of a normal correlation surface, it has frequently been assumed that such a surface is prerequisite to the sound use of the coefficient, but this is not at all true.

Having the means at hand of estimating a second variable, knowing a first, it is desirable to ascertain the probable error of such determinations. Obviously if arrays are homoscedastic, the standard deviation of any array is the standard error of any single estimate.

$$\sigma_{2.1} = \sigma_2\sqrt{1 - r^2} = \sigma_2 k \quad \text{(The standard deviation of an array or the standard error of estimate of a second variable, knowing the first)} \ldots [86]$$

$$\sigma_{1.2} = \sigma_1\sqrt{1 - r^2} = \sigma_1 k$$

The quantity k of the above equations is defined in the next paragraph.

With the data of Chart XX in hand, $\sigma_{2.1} = 2.208\sqrt{1 - (.3888)^2} = 2.034$. That is to say, that if the correlation between height of mid-parent and offspring is .3888 and if the standard deviation of heights of offspring is 2.208 inches, then the standard error of estimate of a child's height, determined from the mid-parent height, is 2.034 inches. A guess that the height of every offspring is $68\frac{1}{4}$ inches would have a standard error of 2.204 inches so that the increased accuracy of estimate due to utilizing the correlation of .3888 between mid-parent and offspring reduces the standard error of estimate to 2.034 inches, or about 8 per cent reduction. It is thus seen that no very great improvement in estimate results from a correlation no higher than .3888. The proportionate reduction is given by the factor $\sqrt{1 - r^2}$. This factor measures the lack of relationship between two variables just as r measures presence of relationship. I have elsewhere (Kelley, 1919) described certain of its properties and have termed it a coefficient of alienation. The coefficient of alienation may be interpreted in a positive sense for if a criterion, x_0, correlates to the extent r with a given measure, x_1, and if there exists some other measure, x_2, independent of x_1 but which together with it completely determines x_0, then the correlation between x_0 and x_2 is k. Its immediate determination, having any value of r, is given by

$$k = \sqrt{1 - r^2} \quad \text{(Coefficient of alienation)} \ldots [86\,a]$$

and the calculation may readily be made by the aid of the small alignment chart given in the appendix or the large chart which is a supplement to (Kelley, 1921). To secure an idea of the

improvement of estimate with increase in correlation, the following table is given:

Coefficient of Correlation	Coefficient of Alienation	Coefficient of Correlation	Coefficient of Alienation
r	k	r	k
.00	1.0000	.80	.6000
.10	.9950	.8660	.5000
.30	.9539	.90	.4359
.50	.8660	.95	.3122
.60	.8000	.98	.1990
.70	.7141	.99	.1411
.7071	.7071	1.00	.0000

Notice that a correlation of .866 is necessary before the error of estimate has been reduced a half, and that even with a correlation of .99, the error of estimate is still 1/7 as great as a sheer random guess. It should be obvious from these facts that if individual estimates are to be made, it is necessary that very high correlation be present in order to secure even moderately reliable results.

It is sometimes convenient to work with probable errors instead of standard deviations, in which case we have

$$\text{P. E.}_{1.2} = \text{P. E.}_1 \, k \quad \text{(Probable error of estimate of the second variable, knowing the first)} \ldots\ldots\ldots [86\ b]$$

The calculation of the formula for the probable error of the coefficient of correlation is involved and has several times been given (Sheppard, 1898), (Pearson, 1913, freq.), and is not repeated here, but the formulas upon which it is based have general value. Not only the probable error of the coefficient of correlation, but many other probable errors as well, depend upon certain higher product moments and upon the correlation between product moments. The notation and meaning of product moments may be made clear by certain illustrations.

$$p_{11} = \frac{\Sigma XY}{N}$$

and is a second order product moment,

$$p_{12} = \frac{\Sigma XY^2}{N}$$

and is a third order product moment,

$$p_{34} = \frac{\Sigma X^3 Y^4}{N}$$

and is a seventh order product moment, etc., and in general,

$$p_{qq'} = \frac{\Sigma X^q Y^{q'}}{N}$$

gives a product moment of the $(q + q')$ order around some fixed point. Following Pearson we would use the symbol $\bar{p}_{qq'}$ to represent the same product moment around the mean as origin, but as moments around the mean are the only ones here concerning us, we will drop the superior bar and use p_{qq} in place of $\bar{p}_{qq'}$. The meaning of the notation may be illustrated by a few examples involving familiar constants.

$$p_{10} = \frac{\Sigma xy^0}{N} = \frac{\Sigma x}{N} = 0$$

$$p_{01} = \frac{\Sigma y}{N} = 0$$

$$p_{20} = \frac{\Sigma x^2 y^0}{N} = \frac{\Sigma x^2}{N} = \mu_2 = \sigma_1^2$$

$$p_{02} = \mu'_2 = \sigma_2^2 \quad \text{(The prime designating the second variable)}$$

$$p_{11} = \frac{\Sigma xy}{N} = r_{12}\sigma_1\sigma_2.$$

Section 49. Standard Deviations and Correlations of Various Constants

The standard error of any product moment is given by the equation (Pearson, 1913, freq.),

$$N\sigma^2_{p_{q,q'}} = p_{2q,2q'} - p^2_{q,q'} + q^2\, p_{20}\, p^2_{q-1,q'} + q'^2\, p_{02}\, p^2_{q,q'-1} + 2\,qq'p_{11}\, p_{q-1,q'}\, p_{q,q-1} - 2\,q\,p_{q+1,q'}\, p_{q-1,q'} - 2\,q'p_{q,q'+1}\, p_{q,q'-1}$$

(Standard error of any product moment from the means) . . .[105]

The correlation between any two product moments is given by

$$N\sigma_{p_{q,q'}}\,\sigma_{p_{u,u'}}\,r_{p_{q,q'}p_{u,u'}} = p_{q+u,q'+u'} - p_{q,q'}\,p_{u,u'} + qu\,p_{20}\,p_{q-1,q'}\,p_{u-1,u'} + q'u'p_{02}\,p_{q,q'-1}\,p_{u,u'-1} + qu'p_{11}\,p_{q-1,q'}\,p_{u,u'-1} + q'u\,p_{11}\,p_{q,q'-1}\,p_{u-1,u'} - u\,p_{q+1,q'}\,p_{u-1,u'} - u'\,p_{q,q'+1}\,p_{u,u'-1} - q\,p_{u+1,u'}\,p_{q-1,q'} - q'p_{u,u'+1}\,p_{q,q'-1}$$

(Correlation between any two product moments taken from the means) .[106]

These two equations provide the basic relationships which lead to the following special probable errors and correlations.

As will be noted the formulas greatly simplify if homoscedasticity and rectilinearity are assumed, and simplify still further if normality of correlation surface is assumed.

Standard error of the mean,

$$\sigma_M = \frac{\sigma}{\sqrt{N}} \quad \text{.....[29]}$$

Standard error of the standard deviation,

$$\sigma_\sigma = \frac{\sigma}{\sqrt{2N}} \quad \text{(Assuming a mesokurtic distribution)..[32 } a\text{]}$$

Standard error of the regression coefficient,

$$\sigma_{b_{12}} = \frac{\sigma_1}{\sigma_2}\sqrt{\frac{1 - r^2}{N}} = \frac{\sigma_1 k}{\sigma_2\sqrt{N}} \quad \text{(Assuming homoclisy and rectilinearity).....[107]}$$

Standard error of the correlation coefficient,

$$\frac{\sigma^2 r_{xy}}{r^2_{xy}} = \frac{1}{N}\left(\frac{p_{22} - p^2_{11}}{p^2_{11}} + \frac{p_{40} - p^2_{20}}{4\,p^2_{20}} + \frac{p_{04} - p^2_{02}}{4\,p^2_{02}} + \frac{p_{22} - p_{20}\,p_{02}}{2\,p_{20}\,p_{02}} - \frac{p_{31} - p_{11}\,p_{20}}{p_{11}\,p_{20}} - \frac{p_{13} - p_{11}\,p_{02}}{p_{11}\,p_{02}}\right)$$

(No assumptions except that [error/N] to second and higher powers are negligible in comparison with first powers)....[108]

This complete formula was first given by Sheppard (1898)

$$\sigma_r = \frac{k^2}{\sqrt{N}}\left(1 - \frac{1}{4}(\beta_2 + \beta'_2 - 6)\frac{r^2}{k^2}\right)^{\frac{1}{2}}$$

(Assuming rectilinearity of regression. This assumption carries with it the necessity of equal kurtosis, if arrays are homoscedastic) .....[108 *a*]

This formula, as well as others in this section, is given by Pearson (1913, freq.). The constants, β_2 and β'_2, are the β_2's for the two distributions.

$$\sigma_r = \frac{k^2}{\sqrt{N}} \quad \text{(From preceding formula, assuming mesokurtosis, in addition).....[108 } b\text{]}$$

This standard error was first derived by Filon and Pearson (1898), upon the assumption of normality, but note that the formula is in fact more general than this. Also note that if r is high and the kurtosis small, the formula gives too small a value; and that if the correlation and kurtosis are high, the formula gives too large a value.

$$\sigma_r = \frac{1 - \rho^2}{\sqrt{N-1}}\left(1 + \frac{11\,\rho^2}{4\,(N-1)}\right) \quad \text{(Standard error of } r \text{ to a second approximation).....[108 } c\text{]}$$

In the derivation of this formula, squares of the magnitudes [error/N] were kept and normality of correlation surface assumed. (Soper, 1913.) The magnitude ρ is the true correlation and for such small populations as this formula is intended it may lead to substantial error to use r, the obtained correlation, in its place. This is particularly true if r is very large. However, the use of r in place of the unknown value, ρ, if $r < .95$ and calculation of the standard error of r by the above formula in case $N < 25$, should give better results than formulas [108 *a*] or [108 *b*]. If formulas [108 *a*] or [108 *b*] are used for these small populations an improved result may be expected by multiplying the standard error given by them by

$$[1 + (1 + 5.5\, r^2)/(2\, N)] \quad \text{.....[108 } d\text{]}$$

As a practical matter, r determined from samples < 5 may be considered meaningless and nearly so if determined from samples < 7.

Standard error of the constant term $(M_1 - b_{12}M_2)$ of the regression equation. Let $c = (M_1 - b_{12}M_2)$. Then,

$$\sigma_c = \sigma_{b12}\sqrt{M^2{}_2 + \sigma^2{}_2} \quad \text{(Assuming homoclisy and rectilinearity) .....[109]}$$

Standard error of the estimated mean of an array, $\bar{y}_x$ (the mean y score of the x-array).

$$\sigma_{\bar{y}_x} = \frac{\sigma_2\, k}{\sqrt{N}}\sqrt{1 + \frac{x^2}{\sigma^2{}_1}} \quad \text{(Assuming homoclisy and rectilinearity) .....[110]}$$

Note the decrease in the accuracy of the means of the arrays as we go further and further from the mean of the total distribution. A further important consequence of this equation is that for certain situations it gives the standard error of the mean of a total population [see formula 111] since the estimated mean of the array for $x = 0$ is the mean of the total y-distribution.

$$\sigma_{M_2} = \frac{\sigma^2\, k}{\sqrt{N}} \quad \text{(Standard error of a second mean in case a first mean is known with zero error, and in case the correlation between the two series of measures is } r\text{) .... [111]}$$

Certain correlations between the constants of a correlation surface are at times needed. Let n_s = the frequency in row

s; n_s' in column s'; and n_{ss}' in the compartment or cell given by the intersection of the s row and the s' column. Then,

$$\sigma^2 n_{ss'} = n_{ss'}\left(1 - \frac{n_{ss'}}{N}\right) \quad \text{[112]}$$

$$\sigma n_{ss'}\,\sigma n_{tt'}\,r n_{ss'}\,n_{tt'} = -\frac{n_{ss'}\,n_{tt'}}{N} \quad \text{[113]}$$

$$\sigma n_s\,\sigma n_{s'}\,r n_s n_{s'} = n_{ss'} - \frac{n_s\,n_{s'}}{N} \quad \text{[114]}$$

$$\sigma n_s\,\sigma n_{ts'}\,r n_s n_{ts'} = -\frac{n_s n_{ts'}}{N} \quad \text{[115]}$$

$$\sigma n_s\,\sigma n_{ss'}\,r n_s n_{ss'} = n_{ss'}\left(1 - \frac{n_s}{N}\right) \quad \text{[116]}$$

$$\sigma_{M_1}\,\sigma n_{ss'}\,r_{M_1 n_{ss'}} = \frac{n_{ss'}}{N}\,x_s$$ (Correlation between the mean and the frequency of a cell) [117]

$$r_{M_1\,M_2} = r_{12}$$ (Correlation between means) [118]

$$r_{\sigma_1\sigma_2} = \frac{p_{22} - p_{20}\,p_{02}}{\sqrt{p_{40} - p^2{}_{20}}\sqrt{p_{04} - p^2{}_{02}}} = \frac{p_{22} - \mu_2\mu'_2}{\sqrt{\mu_4 - \mu_2{}^2}\sqrt{\mu'_4 - \mu'_2{}^2}}$$

(Correlation between standard deviations) [119]

$$\frac{p_{22}}{\mu_2\mu'_2} - 1 = r^2(\beta_2 - 1) = r^2(\beta'_2 - 1)$$

(Assumption that both distributions are homoscedastic and regressions rectilinear) [120]

Thus,

$r_{\sigma_1\sigma_2} = r^2$ (Assumption of rectilinearity, homoscedasticity and equal kurtosis) [121]

$p_{13} = r\,\sigma_1\,\sigma_2{}^3\,\beta'_2$ (Assumption of rectilinearity) [122]

$$r_{M_1\sigma_1} = r_{M_1\mu_2} = \sqrt{\frac{\beta_1}{\beta_2 - 1}}$$ (No assumptions) [123]

If $\beta_1 = 0$, then $r_{M_1\sigma_2} = 0$ [123 *a*]

$$r_{rM_1} = \frac{r(\sqrt{\beta_1} - r\sqrt{\beta'_1})}{2k^2}$$ (Assuming rectilinearity and mesokurtosis) [124]

$r_{rM_1} = 0$ (Assuming rectilinearity mesokurtosis and homoclisy) [124 *a*]

$$r_{r\sigma_1} = \frac{r}{2\,\sigma_r\sqrt{N}}(\sqrt{\beta_2 - 1} - r^2\sqrt{\beta'_2 - 1})$$

(Assuming rectilinearity and homoscedasticity) [125]

$$r_{r\sigma_1} = \frac{r}{\sqrt{2}}$$ (Assuming mesokurtosis in addition to above) [125 *a*]

Let $c = (M_1 - b_{12}M_2)$. Then

$$r_{cb_{12}} = -\frac{M_2\sigma^2{}_1k_2}{N\sigma^2{}_2\sigma_c\sigma_{b_{12}}} \quad \text{(Assuming rectilinearity and homoclisy)} \ldots [126]$$

Let $\beta_{12.3}$, $\beta_{13.2}$, etc., be defined as in Section 80. Then,

$$r_{\sigma_1 r_{23}} = \frac{1}{\sqrt{2}}(r_{12}\beta_{13.2} + r_{13}\beta_{12.3}) \quad \text{(Assumption of normality)} \ldots . [127]$$

$$r_{r_{12}r_{34}} = \tfrac{1}{2}(\beta_{31.2}\beta_{24.3} + \beta_{14.3}\beta_{32.1} + \beta_{13.4}\beta_{42.1} + \beta_{41.2}\beta_{23.4})$$
(Assumption of normality) . . . . . [128]

$$r_{r_{12}r_{13}} = r_{23} - \frac{r_{12}r_{13}(k^2{}_{23} - r^2{}_{12} - r^2{}_{13} + 2\,r_{12}r_{13}r_{23})}{2\,k^2{}_{12}k^2{}_{13}}$$
(Assumption of normality) . . . . . [129]

The last three equations were first given by Filon and Pearson (1898). Formulas for a number of the preceding standard errors and correlations, not involving the assumption of normality, are given by Isserlis (1916). He also gives reduction formulas for higher product moments, such for example as for p_{xyz^2}.

Section 50. Formulas for the Calculation of the Product-Moment Coefficient of Correlation

There are a number of useful variations of form in the product-moment formula. The equivalence of all the following statements should be immediately recognized by the student:

(*a*) $r_{12} = \dfrac{\Sigma z_1z_2}{N}$, in which $z_1 = \dfrac{x_1}{\sigma_1}$, and $z_2 = \dfrac{x_2}{\sigma_2}$

(*b*) $r_{12} = \dfrac{\Sigma x_1x_2}{N\sigma_1\sigma_2}$

(*c*) $r_{12} = \dfrac{\Sigma xy}{N\sigma_1\sigma_2}$

(*d*) $\Sigma xy = Nr_{12}\sigma_1\sigma_2$

(*e*) $p_{11} = r_{12}\sigma_1\sigma_2$, or $r_{12} = \dfrac{p_{11}}{\sigma_1\sigma_2}$

(*f*) $r_{12} = b_{12}\dfrac{\sigma_2}{\sigma_1} = b_{21}\dfrac{\sigma_1}{\sigma_2}$.

(Pearson product-moment coefficient of correlation) [90]

In case a table of squares is employed it is simpler to work with sums and differences than with products: Let d = the difference between two deviations, each taken from its mean. We have

$$\sigma^2d = \frac{\Sigma(x-y)^2}{N} = \frac{\Sigma x^2 - 2\Sigma xy + \Sigma y^2}{N} = \sigma^2{}_1 + \sigma^2{}_2 - 2\,r\sigma_1\sigma_2$$

or,

$$r = \frac{\sigma^2_1 + \sigma^2_2 - \sigma^2_d}{2\,\sigma_1\sigma_2}$$ (Difference formula for r, based upon deviations from means).....[130]

in case x and y are equally variable, so that $\sigma_1 = \sigma_2$, we have,

$$r = 1 - \frac{\sigma^2_d}{2\,\sigma^2}$$ (Difference formula for r in case of equal variability, based upon deviations from means)..[131]

Utilizing the usual relationship between a standard deviation around a mean and that around an arbitrary origin we may express the last two equations in terms of gross scores. Let Σ_1 = the standard deviation of the gross scores X around the origin, $X = 0$; Σ_2 that of the Y's, and Σ_d that of the quantities $(X - Y)$, and let M_1 and M_2 stand for the means, then the following formulas are easily derived from the preceding two.

$$r = \frac{\Sigma^2_1 + \Sigma^2_2 - 2\,M_1\,M_2 - \Sigma^2_d}{2\,\sqrt{\Sigma^2_1 - M^2_1}\,\sqrt{\Sigma^2_2 - M^2_2}}$$ (Difference formula for r based upon gross scores).....[132]

In case the means, and standard deviations, are equal, — such a case as would arise if two similar forms of a test are correlated, the formula becomes

$$r = 1 - \frac{\Sigma^2_d}{2\,(\Sigma^2 - M^2)}$$ (Difference formula for r based upon gross scores and in case means and standard deviations are equal).....[133]

The difference formula based upon gross scores may be transformed into one involving summations instead of averages. Let $S_1 = N\,\Sigma^2_1$, $S_2 = N\,\Sigma^2_2$, $S_d = N\,\Sigma^2_d$, $\Sigma X = NM_1$, $\Sigma Y = NM_2$. Then, we have,

$$r = \frac{\frac{N}{2}\,(S_1 + S_2 - S_d) - (\Sigma\,X)\,(\Sigma\,Y)}{\sqrt{NS_1 - (\Sigma\,X)^2}\,\sqrt{NS_2 - (\Sigma\,Y)^2}}$$ (Difference formula for r based upon sums of gross scores).....[134]

Formulas such as [134] involving gross scores only are advantageous in that they lend themselves readily to mechanical and routine calculation. The numerical figures involved frequently become large but this is not much of a handicap, if a table of squares is used, and if an adding machine is available.

Formulas similar to certain of the preceding, based upon sums instead of differences, are as follows: Let σ_s stand for the standard deviation of the sums of the deviations from the mean $(x + y)$, and Σ_s for the standard deviation from zero

of the sums of gross scores $(X + Y)$, and let other symbols be as above, then

$$r = \frac{\sigma^2_s - \sigma^2_1 - \sigma^2_2}{2\,\sigma_1\sigma_2}$$ (Sum formula for r based upon deviations from means).....[135]

$$r = \frac{\sigma^2_s}{2\,\sigma^2} - 1$$ (Sum formula for r based upon deviations from means in case of equal variability).....[136]

Eliminating σ^2 from formulas [131] and [136] gives

$$r = \frac{\sigma^2_s - \sigma^2_d}{\sigma^2_s + \sigma^2_d}$$ (Sum and difference formula for r based upon deviations from means in case of equal variability) . [137]

If gross scores are used and if means, and standard deviations, are equal, formula [137] may be transformed into the following:

$$r = \frac{\Sigma^2_s - \Sigma^2_d - 4\,M^2}{\Sigma^2_s + \Sigma^2_d - 4\,M^2}$$ (Sum and difference formula for r based upon gross scores in case means, and standard deviations, are equal).....[138]

A general formula based upon the standard deviations of sums may be readily derived and is sometimes useful, as is also one based upon summations of sums.

In general; if, for a given problem, certain relationships are known to hold ahead of calculation, such, for example, as equal means, equal standard deviations, proportionate means, proportionate standard deviations, means or standard deviations having known values, etc., a simpler formula than the general one may be derived. If inexperienced help is doing the work, a mechanical routine method not involving such mental operations as multiplying three times seven, but rather such operations as copying 197244 and adding on an adding machine, is serviceable. If multiplication as high as twelve times twelve, and good judgment in selecting approximate means can be counted upon, the method used upon Galton's data is probably the most expeditious.

Section 51. The Interpretation of Regression Coefficients

The derivation of the correlation coefficient shows it to be the regression coefficient in the case of standard measures. The regression coefficient is statistically the more fundamental and in all actual problems involving the estimate of one variable knowing a second, the regression coefficient and not the correlation coefficient is the essential measure. A wider use of

regression coefficients in place of correlation coefficients would lead to a more accurate and detailed understanding of the situations portrayed. We may illustrate this by the data of Chart XXI, but will first need to know the standard error of a difference. This is readily derived. Let d equal the difference between two measures X and Y, whose means are M_1 and M_2 and let x and y be defined by the equations $x = X - M_1$, $y = Y - M_2$, then

$$d = X - Y = (x - y) + (M_1 - M_2)$$

If any constant is added to or subtracted from d, the standard deviation around the mean is not altered so that

$$\sigma_d = \sigma_{(d + M_2 - M_1)}$$

and since

$$d + M_2 - M_1 = x - y$$

we have

$$\sigma_d = \sigma_{(x - y)}$$

but $\sigma_{(x-y)}$ is simply σ_d of formula [130]. Solving [130] for σ_d we have

$$\sigma_d = \sqrt{\sigma^2_1 + \sigma^2_2 - 2\, r_{12}\, \sigma_1 \sigma_2}$$ (Standard error of the difference between 2 correlated measures). . . . [139]

in which σ_1 is the standard error of the first measure, σ_2 of the second measure, and r_{12} is the correlation between the two measures. In case the measures are not correlated we have

$$\sigma_d = \sqrt{\sigma^2_1 + \sigma^2_2}$$ (Standard error of the difference between two independent measures). . . . . [140]

The constants calculated from this chart, including the correlation ratio η and the test for linearity ζ, described in Section 68, are as given in Table XXXVI, in which variable one is the percentage of men voting for Thompson, and variable two, the percentage of women.

TABLE XXXVI

			Standard errors of			
$M_1 = 60.768$	$M_2 = 60.558$		M_1, .374	M_2, .441		
$\sigma_1 = 14.707$	$\sigma_2 = 17.354$		σ_1, .264	σ_2, .312		
$b_{12} = .73527$	$b_{21} = 1.02377$		b_{12}, .0107	b_{21}, .0149		
	$r_{12} = .86761$			r_{12}, .0063		
$\eta_{12} = .86942$	$\eta_{21} = .87112$		η_{12}, .0061	η_{21}, .0062		
$\zeta_{12} = \eta^2_{12} - r^2_{12} = .00611$	$\zeta_{21} = \eta^2_{21} - r^2_{12} = .00314$		ζ_{12}, .0040	ζ_{21}, .0028		

The method of calculating the standard error of ζ is given later, but since its probable error is nearly as large as itself, rectilinearity is shown to be a sound assumption. Let us, therefore, consider the other constants and attempt to answer the following questions:

(1) Is there a sex difference in regard to the mean tendency: that is, is the difference $(M_1 - M_2)$ which equals .210, one fifth of one per cent, a significant difference?

CHART XXI*

CORRELATION BETWEEN SEX AND VOTING TENDENCIES

PERCENTAGE OF MEN VOTING FOR THOMPSON

PERCENTAGE OF WOMEN VOTING FOR THOMPSON	2	7	12	17	22	27	32	37	42	47	52	57	62	67	72	77	82	87	92	97	TOTAL
2			1																		1
7			2	1																	3
12		2	1	3	1	1	1									1					10
17				2	3	2	4					1									12
22			1	1	1	7	8	1	1	1											21
27				1		7	9	8	5	4	1										35
32						2	8	12	15	2	3		1	1	1	1					46
37			1		1	1	3	17	25	24	9	1	2		1						85
42			1			2	2	8	20	18	14	6	4	2							77
47								1	12	32	32	13	6	3		2					101
52						1	1	1	2	23	53	30	19	5	2	4	1				142
57								2		9	26	40	37	15	3		1		1		134
62								1	2	2	13	33	46	36	10	3		2			148
67										4	5	10	52	66	42	8	4		2		193
72									1	1	1	6	20	51	55	25	3				163
77											3	1	3	30	59	59	8				163
82												2	2	13	35	57	24	6			139
87											1		3	1	6	11	23	11			56
92																4	4	2	4		14
97																		1	2		3
TOTAL		2	7	8	6	23	36	51	83	120	161	143	195	223	214	175	68	22	9		1546

M_2

M_1

* Correlation of percentage of men's votes cast for Thompson (abscissa) and percentage of women's votes cast for Thompson (ordinate) in 1546 precincts in the Chicago municipal election of April 6, 1915. Percentages are of votes cast for the two leading candidates only. Class intervals run from 4.5010 to 9.5000, etc., per cents. The middle of the class intervals are 7.0005, 12.0005, 17.0005, etc. The .0005 has been dropped in the calculations, and the class symbols are given as 7, 12, 17, etc. The number of votes per precinct did not differ greatly and ran about 400 per precinct, about 35 per cent being votes of women. The data were gathered from official returns by Professor J. W. Canning.

(2) Is there a sex difference in regard to the variability of mean precinct votes: that is, is $\sigma_2 - \sigma_1 = 2.647$ a significant difference?

(3) Is there a sex difference in regression of mean precinct votes: that is, is $b_{21} - b_{12} = .28850$ a significant difference? We can answer these questions by using formula [139] if we know (1) the correlation between means, (2) that between standard deviations, (3) that between regression coefficients. By formula [118]

$$r_{M_1M_2} = r_{12} = .8676$$

by formula [12]

$$r_{\sigma_1\sigma_2} = r_{12}^2 = .7527$$

We have no formula for the direct calculation of the correlation between b's, but we do not need one. If the difference $b_{12} - b_{21}$ is significant, then the quotient, b_{12}/b_{21}, is significantly different from 1.00, but $b_{12}/b_{21} = \sigma^2_1/\sigma^2_2$. Therefore if b_{12}/b_{21} is significantly different from 1.00, σ_1/σ_2 is also, but if this is so, then the difference $(\sigma_1 - \sigma_2)$ is significant. Accordingly if we prove that there is a significant difference between the two standard deviations, we have with the same certainty proven that there is a significant difference in the two regressions.

Letting σ_d stand for the difference of the measure under discussion, we have

$$M_1 - M_2 = .210$$

$$\sigma_d = \sqrt{(.374)^2 + (.441)^2 - 2\,(.8676)\,(.374)\,(.441)} = .219$$

$$\sigma_2 - \sigma_1 = 2.647$$

$$\sigma_d = \sqrt{(.264)^2 + (.312)^2 - 2\,(.7527)\,(.264)\,(.312)} = .207$$

As the standard error of the difference between the means is equal to the difference, we cannot conclude that the difference is significant, but as the standard error of the difference between the standard deviations is but 1/12 of the difference, the point is definitely established that there is a sex difference resulting in difference in the standard deviations and in the regressions. In other words, on the average, throughout the city, men and women voted for Thompson to about the same extent, but judging by the precincts, the women tended to vote in blocks to a greater extent than men. If the precinct was a "Thompson precinct" the majority given to Thompson by the women was greater than that given by the men, and if it was an "anti-

Thompson precinct," the majority against Thompson given by the women was greater than that given by the men. One precinct in particular is a notable exception. This is the one recorded in row 12 and column 77 of Chart XXI. There existed in this precinct a very strong anti-Thompson women's organization, with the result that though 77 per cent of the men voted for Thompson, only 12 per cent of the women did so. The two regression lines involved are drawn and the constants given in detail in order to point the significance of regression lines. That there is a correlation between the votes of men and women is of quite secondary interest to the fact that there is a wide difference in the regressions of the two sexes. The interpretation of the correlation table given hinges upon the slopes of regression lines in a much more fundamental sense than upon the value of the correlation.

Section 52. Product-Moment Correlation of Non-Rectilinear Data

We will now consider a problem involving the calculation of a Pearson product-moment coefficient of correlation from non-rectilinear data. I am indebted to Mr. H. A. Richmond for the accompanying problem and data. Each entry in Table XXXVII is for a single state, except the starred entry which is for the District of Columbia. From considerations altogether outside the data it seems appropriate to consider the District of Columbia data not to be homogeneous with the rest, and they are accordingly omitted from calculations.

Chart XXII

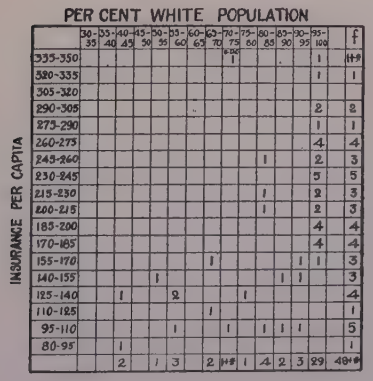


Chart XXIII

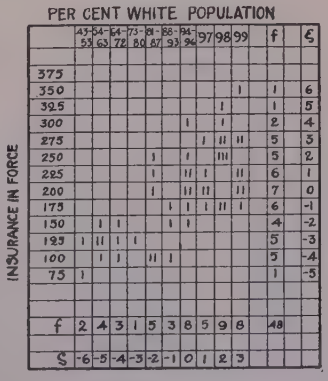


TABLE XXXVII

Per Cent White Population	Per Capita Insurance in Force	Per Cent White Population	Per Capita Insurance in Force
99	341	95	304
99	285	95	251
99	270	95	237
99	219	94	140
99	192	93	103
99	190	90	167
99	170	88	142
99 —	224	87	105
98	321	84	254
98	290	83	207
98	272	82	227
98	269	82	101
98	253	78	133
98	244	*71	*347
98	241	71 —	96
98	182	68	121
98	171	67	158
97	272	58	133
97	234	57	105
97	204	56	126
97	197	54	147
97	182	44	132
96	237	43	84
96	202		
96	190		
96	176		

Let X stand for the per capita insurance in force, and Y for the per cent population, then calculation gives

$$r_{12} = .6430$$

Corrected for fineness of grouping error

$$\eta_{12} = .7955$$

$$\eta_{12} = .7310$$

$$\eta_{21} = .8019$$

Corrected, $\eta_{21} = .7394$

(Calculation given in Section 68)

$$\zeta_{12} = \eta^2_{12} - r^2_{12} = (.7955)^2 - (.6430)^2 = .2193$$

$$\sigma_\zeta = .1202 \quad \text{(Calculation by formula.....[197])}$$

$$\frac{\zeta_{12}}{\sigma_\zeta} = 1.823$$

so that (from Table K–W), the chances are 34 in 1000 that the true regression is rectilinear. The small population makes it impossible to prove the appropriateness of a certain regression line, rectilinear or otherwise, but with only one chance in 30 of the regression being rectilinear, we will proceed on the

assumption that it is definitely non-rectilinear. Since the populations in the successive arrays are very small, the regression line following all the chance fluctuations of the means of the arrays leads to a measure of correlation which is too large to represent the truth. Accordingly .64 is too small and .80 too large, and the true regression is neither a straight line nor one following all the means of the arrays. A value in the neighborhood of .7394 is more trustworthy than either of these. As an empirical procedure, which will result in a more reasonable regression line, and a measure of correlation between .64 and .80, we may use a coarser and coarser grouping of percentages as the data deviate more and more from the mode, assign interval values to grouped data, and calculate a Pearson product-moment coefficient as shown in Chart XXIII. Percentage scores are transformed into auxiliary scores according to the following table:

Per Cent of White Population as Follows . . .	43 to 53	54 to 63	64 to 72	73 to 80	81 to 87	88 to 94	94 to 96	97	98	99
Assign Following Scores	1	2	3	4	5	6	7	8	9	10

This transformation scheme is empirical but it should be noted that it has not been so drawn up as to capitalize chance fluctuations, thus giving a spuriously high measure of correlation. We are not endeavoring to secure a high measure of correlation such, for example, as the raw correlation ratio, but rather a reasonable measure; and second, we desire a procedure which permits estimating one variable, knowing the second, which the correlation ratio method does not permit. We may judge of the excellence of our transformation scheme by the approach of the resulting product-moment coefficient of correlation to the mean of the values of the two corrected correlation ratios $(.7310 + .7394)/2 = .7352$. With this auxiliary score which bears a 1 to 1 relation with percentage of white population, the regression is practically rectilinear. The means of the arrays vary from a position on a straight line only to a degree which we may reasonably attribute to chance. Since there is a 1 to 1 relation between the auxiliary variable

and per cent of white population, an estimation of the auxiliary variable is equivalent to an estimate of the per cent of white population. The Pearson product-moment coefficient of correlation found between the auxiliary score and insurance in force is .7146 which, though it is not quite = .7352, the most reasonable value, is certainly an improvement upon either the straight correlation coefficient or the raw correlation ratio.

In addition to enabling an estimate of one variable from a second, and to providing a reasonable measure of correlation, a reduction of one variable so as to yield a rectilinear regression with a second makes possible an investigation of multiple correlation tendencies which otherwise would be very laborious or altogether impossible.

If we have three variables, X_0, X_1, X_2, and desire to know all the interrelations, we require information as to six regression lines which we may call l_{01}, l_{10}, l_{02}, l_{20}, l_{12}, l_{21}. Let us suppose that the correlation table involving variables 0 and 1, shows 2 rectilinear regressions, l_{01} and l_{10}, and that the regression l_{02} is curvilinear, and that the nature of the others has not been determined. Let us suppose that a simple transformation of X_2 scores into auxiliary X_2' scores results in a rectilinear l_{02} regression line. Then as proven by Isserlis (1914), the additional regression lines l_{20}, l_{12}, and l_{21} are also rectilinear. The proposition may be stated in the words of Isserlis, who uses the word "linear" as we have used rectilinear: "We may conclude then that in general the linearity of any three of the six regression lines involves that of the remaining three." . . . (Isserlis' theorem.)

Obviously the principle can be extended to any number of variables. Let X_0 be the dependent variable or the criterion, and let X_1, X_2, $X_3 \ldots X_n$ be independent variables which are combined into a single score for the purpose of estimating the criterion. Then, if each independent variable showing curvilinear regression with X_0 is transformed into auxiliary scores having rectilinear regression, not only every correlation with the criterion but every intercorrelation between the independent variables as well will be rectilinear. For example, given the four variables X_0, X_1, X_2, X_3. Let us suppose that none of the regressions are rectilinear. In this case the first

investigation to make would be to see if a simple transformation of X_0 may not result in making all the regressions involving X_0 rectilinear. If no such transformation is possible, we may transform the scores of the independent variables. We have the curvilinear regression lines l_{01}, l_{10}; l_{02}, l_{20}; l_{03}, l_{30}; l_{12}, l_{21}; l_{13}, l_{31}; l_{23}, l_{32}. Probably a transformation of some one of the independent variables can be made so that both regression lines involving it and the criterion, that is l_{01}, l_{10}, or l_{02}, l_{20}, or l_{03}, l_{30}, become rectilinear. This is probably always possible in case of single valued functions. Rietz (1919) has shown the impossibility of accomplishing this in the case of multiple valued functions. Let us then so transform X_1, X_2, and X_3 that the following regression lines, l'_{01}, l'_{10}, l'_{02} and l'_{03} are rectilinear. Since l'_{01}, l'_{10} and l'_{02} are rectilinear, we know, by Isserlis' theorem, that l'_{20}, l'_{12} and l_{21} must also be rectilinear, and since l'_{01}, l'_{10} and l'_{03} are rectilinear, l'_{30}, l'_{13} and l'_{31} are also, and since l'_{02}, l'_{20} and l'_{03} are rectilinear, l'_{23} and l'_{32} are also, completing the list. An extension of the method to n variables shows that for the practical purpose of estimating X_1 scores we may make empirical single valued transformations of the dependent variables, wherever necessary to bring about rectilinear regression, and then proceed to calculate the multiple regression equation as described in the next chapter. Thus for single valued functions a lack of rectilinearity ordinarily constitutes no bar to multiple regression procedure.

We have, to this point, considered the significance of correlation as a measure of mutual implication and as a measure derived from the regression coefficient. This interpretation is to be looked upon as basic in correlation treatment. There are, however, other ways of interpreting it, which may occasionally be of value. Weldon (see Brown 1911) has related the correlation coefficient to the percentage of elements which are common to the two series of measures involved. Suppose standing in trait X depends upon the presence or absence of $A + C$ independent elemental factors, and that standing in Y depends upon the presence of $B + C$ independent elemental factors. The C factors are common to both X and Y. The A factors influence X alone and the B factors, Y alone. Further, suppose each factor is as likely to be present as absent, i.e.,

$p = q = \frac{1}{2}$, and when present, to add one half to the trait score, and when absent, to subtract one half from it. Then $x = A + C$; $y = B + C$; and in the long run, $\Sigma A = \Sigma B = \Sigma C = 0$. Let n_a equal the number of A factors, n_b of B and n_c of C factors, then

$$\sigma_A = \sqrt{n_a pq} = \sqrt{n_a \tfrac{1}{2} \cdot \tfrac{1}{2}} = \tfrac{1}{2}\sqrt{n_a}$$

$$\sigma_B = \tfrac{1}{2}\sqrt{n_b}$$

$$\sigma_C = \tfrac{1}{2}\sqrt{n_c}$$

$$\sigma_{A+B} = \tfrac{1}{2}\sqrt{n_a + n_c}$$

$$\sigma_{B+C} = \tfrac{1}{2}\sqrt{n_b + n_c}$$

$$Nr\sigma_1\sigma_2 = \Sigma xy = \Sigma (A + C)(B + C) = \Sigma AB + \Sigma AC + \Sigma BC + \Sigma C^2$$
$$= \Sigma C^2 = N\sigma^2 c$$

since, by supposition, all the elements are independent, all summations of products equal zero. Accordingly

$$r = \frac{n_c}{\sqrt{n_a + n_c}\ \sqrt{n_b + n_c}}$$

If the number of elements determining the score in X equals the number determining that in Y, $n_a = n_b$ and we have

$$r = \frac{n_c}{n_a + n_c}$$

or, the correlation coefficient is the proportion of elements common to the two traits.

Again, suppose trait X is determined by n_c elements and that trait Y is determined by these plus n_b additional ones, that is, $n_a = 0$, then

$$r = \frac{n_c}{\sqrt{n_c}\ \sqrt{n_b + n_c}}$$

and

$$r^2 = \frac{n_c}{n_b + n_c}$$

or, the square of the correlation coefficient is the proportion of elements determining X which are involved in Y. We of course do not know that traits or scores are due to summations of independent elements, so that these results at best have rather doubtful interpretive value, whereas, the interpretation of correlation in terms of regression never fails. Thomson (1919) and Brown and Thomson (1921) deal very fully with this subject.

It has been assumed that the limits of the coefficient of correlation are -1 and 1. This may easily be proven. Let

$$z_1 = \frac{x}{\sigma_1}, \text{ and } z_2 = \frac{y}{\sigma_2}$$

then $\sigma_{z_1} = 1.00$ and $\sigma_{z_2} = 1.00$

$$(z_1 - z_2)^2 > 0$$

$$\frac{1}{N}\Sigma (z_1 - z_2)^2 = \Sigma z_1^2 + \Sigma z_2^2 - 2\Sigma z_1 z_2 = 1 + 1 - 2r$$

but

$$\Sigma (z_1 - z_2)^2 > 0$$

therefore

$$2(1 - r) > 0 \quad \text{or } r < 1$$

Thus the upper limit of r is $+1$.

$$\Sigma (z_1 + z_2)^2 = 2(1 + r) > 0 \text{ or } r > -1$$

Thus the lower limit of r is -1. Accordingly all values of r lie between -1 and 1.

Section 53. The Rank Method of Calculating Correlation

The product-moment method of calculating correlation may be used when differences in merit are expressed in ranks and not in graded scores. Formula [130] is the most convenient to use in deriving the expression for the coefficient of correlation when ranks are used.

The standard deviation of the ranks in the one trait equals σ_1, and of course equals the standard deviation in the other trait, σ_2, as the number of ranks is the same in the two cases. It should, however, be noted that if scores such as

95 94 90 90 87 85 85 85 81 89 75

are assigned ranks

1 2 $3\frac{1}{2}$ $3\frac{1}{2}$ 5 7 7 7 9 10 11

the standard deviation of these pseudo ranks is not identical with that of ranks 1, 2, 3, 4, 5, 6, 7 8, 9, 10, 11. Only slight error is introduced in case ranks are but occasionally divided between two paired measures, but if there are many individuals all given the same rank decided error is present.

Since the standard deviations are equal the equation becomes, using ρ in place of r as is customary when dealing with ranks:

$$\rho = r = 1 - \frac{\Sigma d^2}{2N\sigma^2}$$

The Σd^2 is to be determined by recording the differences in ranks of the individuals in the two traits, squaring and summing. The common standard deviation, σ, may be found from the number of ranks, which is also N, the population. It is only necessary, therefore, to determine the standard deviation of the series $1, 2, 3, \ldots N$ around its own mean. We have

$$\bar{\mu}_1 = \frac{1 + 2 + 3 + \cdots N}{N} = \frac{N + 1}{2}$$

$$\bar{\mu}_2 = \frac{1 + 4 + 9 \cdots N^2}{N} = \frac{4N^2 + 6N + 2}{12}$$

This value for $\bar{\mu}_2$ may be obtained by first determining the second moment, $\bar{m}_2$, in case the distribution consists of frequencies evenly spread over the class intervals, as indicated in the accompanying figure, instead of being concentrated at the class indexes or mid-points as is the case when measures

of rank position are used. The frequency distribution drawn is represented by the line $y = 1$ and extends from $x = \frac{1}{2}$ up to $x = N + \frac{1}{2}$. The second moment from o of any one rank, let us say the k'th, is k^2, whereas the second moment of the distribution $y = 1$ from $(k - \frac{1}{2})$ to $(k + \frac{1}{2})$ is given by the equation

$$\int_{k-\frac{1}{2}}^{k+\frac{1}{2}} yx^2 dx = \frac{x^3}{3}\Big]_{k-\frac{1}{2}}^{k+\frac{1}{2}} = k^2 + \tfrac{1}{12}$$

The moment of the frequency $y = 1$ corresponding to this k'th rank, $1/N$ of the population, is $1/12$ too large, as is of course the case for every other rank; hence the second moment of the equation $y = 1$ from $x = \frac{1}{2}$ to $x = N + \frac{1}{2}$ will be larger than the desired second moment by

$$\frac{1}{N}\left(\frac{N}{12}\right)$$

That is

$$\bar{m}_2 = \bar{\mu}_2 + \tfrac{1}{12}$$

$$\bar{m}_2 = \frac{1}{N}\int_{\frac{1}{2}}^{N+\frac{1}{2}} yx^2 dx = \frac{4N^2 + 6N + 3}{12}$$

Therefore $$\bar{\mu}_2 = \frac{4N^2 + 6N + 2}{12}$$

$$\mu_2 = \bar{\mu}_2 - \bar{\mu}_1{}^2 = \frac{N^2 - 1}{12}$$ (The second moment of N ranks).....[141]

Finally $$\rho = 1 - \frac{6\Sigma d^2}{N(N^2 - 1)}$$ (Spearman's formula for the coefficient of correlation calculated from ranks) ...[142]

This formula should not be confused with Spearman's foot rule formula for correlation

$$R = 1 - \frac{6\Sigma G}{N^2 - 1}$$ (Spearman's foot rule formula for correlation based upon the sum of the gains in rank).....[143]

which has a large, though, except in the case of zero correlation, not definitely known probable error; does not vary between -1 and $+1$; is not at all comparable in meaning with a product-moment coefficient; and in general has none of the merits except brevity, of the formula based on the squares of differences in rank. The coefficient calculated by formula [142] is usually designated by ρ, but it should be noted that it is identical with r if ranks constitute the scores.

Pearson has shown that if scores in the two traits which are in truth normal in form are assigned ranks and ρ calculated, it will differ slightly from the r obtained directly from the scores. To allow for this discrepancy, ρ's may be turned into r's by the formula,

$$r = 2\sin\frac{\pi}{6}\rho$$ (Pearson's correction to Spearman's ρ)..[144]

That the correction is of small magnitude is shown by the accompanying table:

TABLE XXXVIII

ρ	r	ρ	r
.00	.000	.60	.618
.10	.105	.70	.717
.20	.209	.80	.813
.30	.313	.90	.908
.40	.416	.95	.954
.50	.518	1.00	1.000

The formula for ρ is the best of the rank formulas, but in case scores constitute the basic data there is always some loss in accuracy from warping the data into ranks. The probable error of ρ as determined by Pearson (1907 further) is

$$\text{P. E.}_\rho = .7063 \frac{1 - \rho^2}{\sqrt{N}} \quad \text{(Probable error of } \rho\text{)} \ldots [145]$$

or approximately 5 per cent greater than the probable error of r.

In case one of the variables is given in terms of ranks and the other in terms of variates, we may assign rank values to the variates and use formula [142]. If the grouping in the variate series is coarse, ranks cannot be assigned without losing much of the refinement of the variate data, and if the average of a number of ranks is assigned to all the measures in one class there is a further error if formula [142] is used as this formula presupposes serial ranks from 1 to N.

To obviate these difficulties it is better to calculate the product-moment coefficient of correlation between the ranks on the one hand and the variates on the other. Let us call this ρ', and let r be the correlation if the two series could each be expressed in terms of variates and if they constitute a normal correlation surface. Then Pearson (1914, ext.) has shown that,

$$r = \sqrt{\frac{\pi}{3}}\rho' \quad \text{(To deduce } r \text{ from } \rho', \text{ the product-moment correlation between a variate series and a rank series)} \ldots [146]$$

or

$$r = 1.0233\,\rho'$$

PROBLEMS

1. Plot the correlation table giving the correlation between the Thorndike and Ayres scores in handwriting given in Table XXX, Section 34, and answer the question, "Is the relationship between the two variables rectilinear?" *Ans.* It is.

2. Calculate the correlation between series 1 and 2, between series 1 and 3, and also between series 2 and 3 of the paired practice series given in problem 3, Chapter III.

3. Calculate the standard error of r_{12}, the correlation between series 1 and series 2 (*a*) by formula [108 *b*], (*b*) by formula [108 *a*], (*c*) by formula [108 *c*] and finally, as the most accurate method of all, (*d*) by formula [108 *a*] using in addition [108 *d*].

4. Rank the measures in these three series and calculate the correlations ρ_{12}, ρ_{13} and ρ_{23} by formula [142].

5. Determine for the first two of these three series the regression equation for estimating variable 1 from variable 2 and calculate the standard errors of the two constants, b_{12} and c, involved.

6. In the derivation of b_{21} it was assumed that the regression line passed through the means of the two distributions. Derive the same value as b_{21} without making this assumption.

CHAPTER IX

FUNCTIONS INVOLVING CORRELATED MEASURES

Section 54. Correlations of Sums or Averages

If the basic means and standard deviations of several series of measures and the correlations between series are known, the means, standard deviations and correlation of any weighted average or sum of these measures with a second weighted sum may be determined (Spearman, 1913). Given the several series $(a + b)$ in number, $X_1, X_2, \ldots X_a, X_{a+1}, X_{a+2} \ldots X_{a+b}$, with means $M_1, M_2, \ldots M_{a+b}$, standard deviations $\sigma_1, \sigma_2 \ldots \sigma_{a+b}$, and intercorrelations $r_{12}, r_{13} \ldots r_{1\,(a+b)}, r_{23} \ldots$, let the standard measures for these variables be, as usual,

$$z_1 = \frac{X_1 - M_1}{\sigma_1}, \quad z_2 = \frac{X_2 - M_2}{\sigma_2}, \quad \text{etc.}$$

If a of the measures are combined by adding into a single score, and if the remaining b measures are also combined, the correlation between the two composites is

$$r_{(1+2+\ldots a)\,(1+11+\ldots b)} = \frac{\Sigma\,(z_1 + z_2 + \cdots z_a)\,(z_{\mathrm{I}} + z_{\mathrm{II}} + \cdots z_b)}{\sqrt{\Sigma\,(z_1 + z_2 + \cdots z_a)^2}\,\sqrt{\Sigma\,(z_{\mathrm{I}} + z_{\mathrm{II}} + \cdots z_b)^2}}$$

The product of the two terms in parentheses in the numerator gives a binomial of ab terms each of which is a sum of the sort $\Sigma z_1 z_{\mathrm{I}}$, but

$$\Sigma\, z_1 z_{\mathrm{I}} = N r_{1\mathrm{I}}, \quad \Sigma\, z_1 z_{\mathrm{II}} = N r_{1\mathrm{II}}, \quad \text{etc.}$$

Accordingly the numerator equals $N \overset{ab}{\underset{1}{S}} r_{pQ}$. The symbol $\overset{ab}{\underset{1}{S}}$ stands for a double summation, p taking in turn the values in the series from 1 to a, and Q in turn the values from I to b. The square of the first polynomial in the radical in the de-

nominator gives a polynomial of a^2 terms, a of them being of the sort Σz^2_1 and the balance $(a^2 - a)$ of the sort $\Sigma z_1 z_2$. But

$$\Sigma z^2_1 = N, \qquad \Sigma z^2_2 = N, \qquad \text{etc.}$$
$$\Sigma z_1 z_2 = N r_{12}, \qquad \Sigma z_1 z_3 = N r_{13}, \qquad \text{etc.}$$

Further

$$\Sigma z_1 z_2 = \Sigma z_2 z_1$$

and as both of these occur in the summation, there are but $(a_2 - a)/2$ different product sums involved, though each of these is found twice. Accordingly the magnitude under the first radical equals

$$N \overset{a}{\underset{1}{S}} 1 + N \overset{(a^2 - a)}{\underset{1}{S}} r_{pq}$$

in which $\overset{a}{\underset{1}{S}} 1$ is simply 1 added a times so that $\overset{a}{\underset{1}{S}} 1 = a$: $\overset{(a^2-a)}{\underset{1}{S}} r_{pq}$ is a double summation in which p takes all values from 1 to a, and q all values other than p from 1 to a. Thus again each r occurs twice, once as r_{pq} and once as r_{qp}. But an r with repeated subscript, such as r_{pp}, is not found in the summation. The summation under the second radical is similar in type, so that

$$r_{(1+2+\ldots a)(I+II+\ldots b)} = \frac{\overset{ab}{\underset{1}{S}} r_{pQ}}{\sqrt{a + \overset{(a^2 - a)}{\underset{1}{S}} r_{pq}}\sqrt{b + \overset{(b^2 - b)}{\underset{1}{S}} r_{PQ}}}$$

(Correlation between sums or averages of scores). [147]

The preceding formula may readily be generalized so as to apply when gross weighted scores are combined. Let w_1 be the weight of X_1, w_2 of X_2, etc. Then we desire the correlation between $(w_1X_1 + w_2X_2 + \ldots w_aX_a)$ and $(w_IX_I + w_{II}X_{II} + \ldots w_bX_b)$ which may be represented by the symbol

$$r_{(Sw_pX_p)(Sw_PX_P)}$$

In calculating the correlation, each variable must be expressed as a deviation from its own mean. Accordingly $(w_1M_1 + w_2M_2 + \ldots w_aM_a)$ must be subtracted from the first summation variable. This leaves $(w_1x_1 + w_2x_2 + \ldots w_ax_a)$. Similarly for the second summation variable. Proceeding as before we have in place of Σz^2_1 the expression

$$\Sigma (w_1x_1)^2$$

and in place of $\Sigma z_1 z_2$ the expression

$$\Sigma w_1x_1w_2x_2$$

so that finally we obtain

$$r_{\left(\sum_1^a w_p X_p\right)\left(\sum_1^b w_P X_P\right)} = \frac{\sum_1^{ab} w_p \sigma_p w_Q \sigma_Q r_{pQ}}{\sqrt{\sum_1^a w^2{}_p \sigma^2{}_p + \sum_1^{(a^2-a)} w_p \sigma_p w_q \sigma_q r_{pq}} \sqrt{\sum_1^b w^2{}_P \sigma^2{}_P + \sum_1^{(b^2-b)} w_P \sigma_P w_Q \sigma_Q r_{PQ}}}$$

(Correlation between the sums or averages of weighted scores).....[148]

Note that there is nothing in the derivation to prevent certain of the weights being negative. If the correlation between two series is r, this is not changed when all the measures in the first series are divided by a certain quantity and all those in the second by another. Thus in the preceding, division of the first series by a and of the second by b, leading to averages, will not change the correlation. The formula given is therefore equally applicable whether dealing with sums or with averages.

In case a single score is correlated with the weighted average of a number of others we have a situation represented by one of the two sums having but one item in it. Then the summation $\sum_1^b$ has but a single term and $\sum_1^{b^2-b}$ has no terms. Further, $w_1\sigma_1$ cancels from numerator and denominator of the right hand member. This is the very common situation where one variable, which we may call the criterion and represent by X_0, is taken as a standard and all the others are combined so as to give a high correlation with this one. Under these conditions formula [148] becomes:

$$r_{x_0(\sum w_p X_p)} = \frac{\sum_1^a w_p \sigma_p r_{0p}}{\sqrt{\sum_1^a w^2{}_p \sigma^2{}_p + \sum_1^{a^2-a} w_p \sigma_p w_q \sigma_q r_{qp}}}$$

(Correlation between a criterion and the weighted sum or average of a number of scores).....[149]

Since this formula gives the correlation whatever the σw products, or the effective weights, may be, one may frequently by successive trials hit upon a weighting which gives a fairly satisfactory correlation. If two independent variables are involved and the nominal weight of the first independent variable

is arbitrarily set equal to 1.0 while that of the second is indeterminate and called w we have

$$r_{x_0(X_1+wX_2)} = \frac{\sigma_1 r_{01} + w\sigma_2 r_{02}}{\sqrt{\sigma_1^2 + w^2\sigma_2^2 + 2\, r_{12}\sigma_1 w\sigma_2}} \quad \ldots\ldots\ldots [150]$$

The multiple correlation $r_{x_0(X_1+wX_2)}$ and the weight w are the only unknowns in this equation, so it may be plotted on two axes, w the abscissa and r the ordinate, throwing into clear relief the effect of approximate weightings. Thurstone (1919) has shown the value of this procedure. A plot of the following data will illustrate the falling off in the multiple correlation obtained as w varies from $-.9310$, which is the ratio of the regression coefficients $b_{02.1}/b_{01.2}$. Given $r_{01} = .4$, $r_{02} = -.3$, $r_{12} = .12$, $\sigma_1 = \sigma_2 = 1.0$

If $w =$	$-\infty$	-2.0	-1.5	-1.0	$-.9310$	$-.9$
Then $r_{x_0(X_1+wX_2)} =$	.300	.620	.706	.7826	.7846	.7842

$w =$	$-.8$	$-.5$	.0	1.0	1.333	2.0	∞
$r_{x_0(X_1+wX_2)} =$	.776	.682	.400	.056	.000	$-.074$	$-.300$

Returning to [149], in case all of the series summated or averaged have equal standard deviations and are given equal weight, we have:

$$w_1 = w_2 = \cdots w_a = w$$
$$\sigma_1 = \sigma_2 = \cdots \sigma_2 = \sigma$$
$$\sum_1^a w_p\sigma_p r_{op} = w\sigma \sum_1^a r_{op} = aw\,\sigma\, r_c$$

where r_c is the average correlation of the various series with the criterion x_0.

$$\sum_1^a w^2{}_p\sigma^2{}_p = aw^2\sigma^2$$
$$\sum_1^{a^2-a} w_p\sigma_p w_q\sigma_q r_{pq} = w^2\sigma^2 \sum_1^{a^2-a} r_{pq} = w^2\sigma^2\,(a^2 - a)\, r_i$$

where r_i is the average intercorrelation between the several original series so that, finally,

$$r_{x_0(Sw_pX_p)} = r_{x_0(SX_p)} = \frac{ar_c}{\sqrt{a + (a^2 - a)\, r_i}}$$

or,

$$r_{x_0(Sw_pX_p)} = \frac{r_c}{\sqrt{\frac{1 - r_i}{a} + r_i}}$$

(Correlation between a criterion and the sum or average of a number of equally weighted scores) . . [151]

If the tests are comparable the several correlations with the criterion differ but little and any one of them may be taken as a first approximation to r_c, and the intercorrelations differ but little and any one of them is a first approximation to r_i; also $Sw_pX_p = af$ as defined in the next section (55), so that we have

$$r_{0(af)} = \frac{r_c}{\sqrt{\frac{1 - r_{11}}{a} + r_{11}}}$$

(Correlation between a criterion and the sum or average of a number of equally weighted similar test scores) . . . . . [152]

The effective weight given a test is not w_p, the nominal weight, but $w_p\sigma_p$, the product of the nominal weight and the standard deviation of the scores. Accordingly equally weighted scores are those in which the products of the nominal weights and the standard deviations are equal; that is, if $w_1\sigma_1 = w_2\sigma_2 = w_3\sigma_3 = \cdots$, etc., the X_1, X_2, X_3, etc., series or scores are actually weighted equally. This is the condition that must hold if the immediately preceding formula is to remain true.

Section 55. The Reliability Coefficient

Let us suppose that the scores combined are those of comparable tests of some single function. If the tests are strictly comparable, then in addition to the means, and standard deviations being equal

$$r_{01} = r_{02} = \cdots = r_c$$

and

$$r_{12} = r_{13} = \cdots = r_{23} = \cdots r_{11}$$

the correlation between one form or test and a second similar form. Let us define a "true score" as the average score on an infinite number of strictly comparable tests. Then the correlation between the criterion and such a true score, which can be obtained by letting a of formula [150] become infinite, may be written as

$$r_{0\infty} = \frac{r_{01}}{\sqrt{r_{11}}}$$

(Correlation between a fallible criterion and a true score) . . . [153]

in which ∞ designates the infinite summation. If the reliability coefficient of the criterion r_{00} is known, we have, as the same sort of formula as [153]

$$r_{1\infty} = \frac{r_{01}}{\sqrt{r_{00}}} \quad \text{(Correlation between a true criterion and a fallible score) [153a]}$$

The correlation r_{01} is that between a test and a criterion, and r_{11} is that between two comparable tests and is called a reliability coefficient. That the notation may be entirely clear, the meanings of several symbols as they will be used are here listed. $r_{af,Af}$ is the correlation between the sum, or average, of a measures of a certain sort and A others of the same sort. Capital A is used in the second subscript instead of small a to indicate that the second series of tests (the same in number as in the first series) is different, though similar to the a tests averaged or summated in the first series. Whenever a is greater than one, the f is kept in the subscript, but when a single test is correlated with a single other test, it is dropped, and the subscript designates the variable. Thus $r_{2f,11f}$ means that an average or sum of two forms of the test (or average or sum of two comparable measures of whatever sort they may be) are correlated with the average or sum of two other comparable forms and r_{211} means that one form of the test 2 is correlated with a second similar form of the same test. In this latter case 2 refers to the variable, whereas in the former case (2 f) the 2 refers to the number of forms averaged or summed. The symbol r_{11} represents the correlation between retestings with the same form. If the variable X_1 is a test score the only reason r_{11} does not equal 1.0 is that there is a time interval between the two answers, which an individual gives to the same question. Similarly $r_{af,af}$ means the correlation between average scores upon re-testing with the same a forms.

Certain very specific conditions need to hold before two tests may be considered comparable, and therefore before a correlation between two tests can be considered a reasonable reliability coefficient. In educational and psychological testing the first of two similar tests frequently calls forth a response which is different from the second. The greater familiarity with the form of the test or the difference in interest aroused

may make the second test quite different from the first. This would be especially true if certain elements in the first were so similar to elements in the second as to lead to what may be called a memory transference from the first test to the second. For example, suppose the following questions occur in the first and second tests respectively:

"(*a*) John is taller than James and James is as tall as Joe. Joe is shorter than Jack. How do John and Jack compare in height?"

"(*b*) Bessie is brighter than Bertha, and Bertha is just as bright as Beula. Beula is not quite as bright as Beatrice. Which is the brighter, Beatrice or Bessie?"

One would expect memory transference, and a tendency to solve the second in the same way as the first. We may call such a situation one in which there is a correlation between errors, meaning that, whatever elements of uncertainty or chance operated in the solution of the first question, they would tend to operate in the same manner in the solution of the second. This situation would tend to make r_{1I} too high as a true measure of reliability. There are other, and usually more important, factors which operate in the other direction. Let us suppose the two following questions occur in two forms and that they are intended to be comparable: "(*a*) Who was the first president of the United States?" and "(*b*) Who was the leading batter in the American League in 1920?" Passing over the possibility of some other question than (*a*) in the first test being comparable to (*b*) and some other than (*b*) in the second test being comparable to (*a*), let us consider the comparability of the two questions given. There is certainly no memory transference which would help or hinder in answering (*b*) after having answered (*a*), but the ability to answer (*a*) probably tests special capacity or knowledge which is quite different from that demanded for the correct answering of (*b*). In other words (*a*) and (*b*) are not samplings of the same capacity and two tests made up of questions no more similar than (*a*) and (*b*) can hardly be considered comparable, and as a consequence they would lead to an r_{1I} which would be too small. This is the situation which is the more likely and the more serious as $r_{0\infty}$ in this case becomes too large. The

errors of interpretation due to a too large estimated correlation between a test score and a criterion are probably in general more serious than those due to a too small estimated correlation.

The following rule for the construction of two comparable tests may be laid down: (1) sufficient fore-exercise should be provided to establish an attitude or set, thus lessening the likelihood of the second test being different from the first, due to a new level of familiarity with the mechanical features, etc.; (2) the elements of the first test should be as similar in difficulty and type to those in the second, pair by pair, as possible; but, (3) should not be so identical in word or form as to commonly lead to a memory transfer or correlation between errors.

It is obvious that condition (3) is not met if a test is merely repeated. Only in case the repetition be at so remote a time from the first test that no memory of the earlier response could influence the later would there be no correlation between errors — in fact even were there no conscious memory of the earlier situation there might be a subconscious influence resulting in correlation between the errors. Accordingly the repetition of a test to secure a reliability coefficient is to be deprecated. However, the repetition of a test to secure an upper limit or maximum value above which the true reliability coefficient will not lie may be considered to be a sound procedure.

Spearman (1904 and 1907), who introduced the term "reliability coefficient," used it as here to designate r_{11}, the correlation between comparable tests, and Brown (1911) used the term to mean r_{11}, the correlation between repeated tests. This is an unfortunate vitiating of the Spearman concept. Particularly in view of the fact that a reliability coefficient in the Spearman, and not in the Brown, sense, is the one needed in all the formulas leading to an estimation of true correlation.*

It has been pointed out that the correlation between repeated tests constitutes an upper limit of the reliability coefficient, while the correlation between two forms meeting condition (3), but not fully meeting condition (2), would constitute a lower limit. Should these two correlations lie close together prob-

* The unfortunate use of r_{11} as a reliability coefficient given in Brown (1911) is corrected in the later edition as Brown and Thomson (1921) define r_{11} as here used to be the reliability coefficient.

ably an average of them would constitute a close approximation to the true reliability coefficient. We may expect in most mental and educational test work that the true reliability coefficient will be less than the obtained $r_{1\mathrm{I}}$, and greater than the obtained $r_{1\mathrm{I}}$. The lack of fulfillment of condition (1) for certain age groups and with certain tests probably at times leads to too high a reliability coefficient and at other times to one which is too low.

Section 56. Correction for Attenuation

Let us return to formula [147] and write $\bar{r}_{pQ}$ for the average of all the r_{pQ}'s. Then we have

$$ab\,\bar{r}_{pQ} = \sum_{1}^{ab} r_{pQ}.$$

Similarly

$$(a^2 - a)\,\bar{r}_{pq} = \sum_{1}^{a^2 - a} r_{pq}$$

and

$$(b^2 - b)\,\bar{r}_{PQ} = \sum_{1}^{b^2 - b} r_{PQ}$$

This gives

$$r_{(1+2+\ldots a)\,(\mathrm{I}+\mathrm{II}+\ldots b)} = \frac{ab\,\bar{r}_{pQ}}{\sqrt{a + (a^2 - a)\,\bar{r}_{pq}}\sqrt{b + (b^2 - b)\,\bar{r}_{PQ}}}$$

(Correlation between sums or averages of equally weighted scores) .....[154]

If we make both a and b infinite, we obtain an estimate of the correlation between a true criterion and a true test score, which Spearman calls the value corrected for the attenuation in the raw r_{pQ} value due to chance errors. Let us designate the scores which enter into the criterion as X_1, X_3, X_5, etc., and those entering into the composite test score as X_2, X_4, X_6, etc. Then from [154] we have

$$r_{\infty\,\infty} = \frac{r_{12}}{\sqrt{r_{13}}\sqrt{r_{24}}}$$

(Correlation between a true criterion and true test score, Spearman's formula for correction for attenuation)...[155]

or in the previous notation where r_{12} is the correlation between two different measures, $r_{1\mathrm{I}}$ the reliability coefficient of the first measure, and $r_{2\mathrm{II}}$ of the second, we have

$$r_{\infty\,\infty} = \frac{r_{12}}{\sqrt{r_{1\mathrm{I}}}\sqrt{r_{2\mathrm{II}}}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots [155\,a]$$

The observations as to comparable tests apply equally to the securing of comparable criterion scores. In particular if the criteria are teachers' judgments there may be high correlation between errors in judgments if teachers have discussed certain pupils with each other.

Section 57. Reliability of Averages

Formula [147] for the correlation between sums enables us to determine the reliability of the sum or average of a number of similar tests, knowing the reliability of a single test. If the tests are similar, we may call the successive tests different forms of the same test. Then the standard deviations are equal; if a straight average is taken all weights equal one; and further, if the forms in the $\overset{a}{\underset{1}{S}}$ average are similar to those in the $\overset{b}{\underset{1}{S}}$ average, then every r_{pQ} = every r_{pq} = every r_{PQ} $= r_{1\text{I}}$ – the correlation between one form and a second similar one. Let $r_{af,\,bf}$ be an abridged notation for $r_{\overset{a}{\underset{1}{SX_p}}\ \overset{b}{\underset{1}{SX_p}}}$; that is, for the situation which holds when the scores in both of the summations are upon similar tests or forms. This is the correlation between the average or sum of a forms and the average or sum of b others. It is given by

$$r_{af,\,bf} = \frac{ab\,r_{1\text{I}}}{\sqrt{a + (a^2 - a)\,r_{1\text{I}}}\sqrt{b + (b^2 - b)\,r_{1\text{I}}}}$$

(Correlation between the average score upon a forms and the average upon b others) .....[156]

If a equals b we have:

$$r_{af,\,Af} = \frac{a\,r_{1\text{I}}}{1 + (a - 1)\,r_{1\text{I}}}$$

(The correlation between the average score on a forms of a test and a other similar forms).....[157]

This formula given by Brown (1911) has frequently been called "Brown's formula." It is, however, but a special case of Spearman's earlier formula [147]. If but a single form of a test is available it may be possible to divide it into two comparable halves; for example, one half composed of the

odd and the other half composed of the even exercises, and calculate the reliability coefficient of the half form, $r_{\frac{I}{2}f,\ \frac{I}{II}f}$ or more simply, $r_{\frac{I}{2}\frac{I}{II}}$ and then by formula [157] obtain the reliability coefficient of the single test.

$$r_{II} = \frac{2\, r_{\frac{I}{2}\frac{I}{II}}}{1 + r_{\frac{I}{2}\frac{I}{II}}}$$

(Reliability of a test determined from the scores on the two halves) ..... [158]

A second use to which formula [157] may be put is in the determination of the number of forms required to secure a desired or essential reliability coefficient. Solving for a we obtain

$$a = \frac{r_{af,\ Af}\,(1 - r_{II})}{r_{II}\,(1 - r_{af,\ Af})}$$

(Number of forms required to secure a given reliability $r_{af,\ Af}$) ..[159]

The use of this equation frequently enables one to determine whether it is worth while to attempt to improve a correlation with a criterion by increasing the length of the test. If we have a problem requiring a correlation of not less than .90 with a certain criterion, and not permitting a test program extending over more than two hours, and if we find experimentally that the reliability of a certain 10 minute test is .20 we may determine whether it is of any use continuing with this test. The test cannot, except as a matter of chance, correlate with any criterion to a greater extent than it correlates with a "true" score of the particular function which it measures. Thus if the criterion is the true score in formula [153] then $r_{0\infty}$ becomes $r_{1\infty}$ and r_{01} becomes r_{1I}, so that we have

$$r_{1\infty} = \sqrt{r_{II}}$$

(Correlation between one form of a test and a true score of the function measured by the test)..... [160]

Thus in our present problem $.90 = \sqrt{r_{af,\ Af}}$ or $r_{af,\ Af} = .81$. That is, even if the criterion is no different in its essential nature from that which is measured by the test, it is still necessary to have a test with a reliability of .81 in order to obtain a correlation of .90 with the criterion. Using formula [159] we have

$$a = \frac{.81\ (1 - .20)}{.20\ (1 - .81)} = 17$$

Thus a test at least seventeen times as long as the one with reliability .20 is needed. This would require 170 minutes testing time, which according to conditions laid down is out of the question, so that there is no use continuing with this particular test. This very practical answer is obtained without any knowledge of the criterion or of the test correlation with the criterion.

Formula [152] aids in determining the fitness of a test for a given purpose. Let us suppose that we have three 10 minute tests, the first with reliability .80, the second with reliability .40, and that these two correlate with a criterion to the extent of .30, and that the third test has a reliability of .20 correlating with the criterion to the extent of .24. How much will these correlations be raised by lengthening and thereby making the tests more reliable? Using formula [152] we obtain the accompanying table.

TABLE XXXIX

Length of Test	Time Required in Minutes	Correlation of Scores of Tests of Different Lengths with the Criterion		
		Test X [Reliability .8]	Test Y [Reliability .2]	Test Z [Reliability .2]
¼ of test . . .	2.5	.24	.18	.13
½ of test . . .	5	.27	.24	.17
Single test . . .	10	.30	.30	.24
Sum of 2 tests .	20	.32	.36	.30
Sum of 3 tests .	30	.32	.39	.34
Sum of 5 tests .	50	.33	.42	.39
Sum of 10 tests .	100	.33	.44	.44
Sum of 20 tests .	200	.33	.46	.48
Sum of ∞ tests .		.34	.47	.52

From this table it is apparent that the relative excellence of a test in comparison with others is a matter of reliability, correlation with the criterion, and possibility of increasing or decreasing the length of the test without changing its essential nature. If the three tests can be lengthened or shortened without changing their essential nature then 2.5 or 5 minutes testing with test X would yield a higher correlation with the criterion than the same amount of time with either test Y or Z. Thus if the testing time is less than ten minutes test X is

the most valuable. If the testing time lies between ten minutes and 100 minutes test Y is the most valuable, and if the testing time is over 100 minutes test Z is the most significant. The principle here involved may frequently be used in making the original selection of one or more tests and before correlation with a criterion is known. If the testing time is of necessity brief, give prime consideration to reliability of test; and if the testing time is long, give prime consideration to "validity," to use a term recently employed in psychological literature, i.e., to the accuracy and detail with which the test parallels the criterion function, and but secondary attention to the reliability of the test. If the reliability of the criterion is known the correlations of the tests with a true criterion may be obtained from the coefficients in Table XXXIX by dividing each by the square root of the reliability coefficient of the criterion. The resulting table will show even more strikingly than does Table XXXIX the relative merits of the three tests.

Section 58. The Probable Error of a Coefficient Corrected for Attenuation

The student should carefully note that the coefficient of correlation obtained by the use of the Spearman formula for correction for attenuation should never be used for the estimation of one actual measure from a second. This "corrected" coefficient is a promise of the correlation that one might expect to find between the variables if one had perfectly reliable measures. To use this corrected coefficient in a regression equation would lead to a less close fit of the regression line and to a larger standard error of estimate of the criterion, knowing the independent variable, than occurs when the "raw" correlation coefficient is used. The corrected coefficient of correlation is mainly of value in theoretical discussions and in serving this purpose its divergence from 1.00 is usually material. The derivation of a formula for the standard error of a corrected coefficient is as follows, in which the subscripts have the meanings stated at the beginning of Section 56.

$$r_{\infty\infty} = \frac{r_{12}}{\sqrt{r_{13}}\sqrt{r_{24}}}$$

Taking logarithmic differentials, we have

$$\frac{dr_{\infty\infty}}{r_{\infty\infty}} = \frac{dr_{12}}{r_{12}} - \frac{dr_{13}}{2r_{13}} - \frac{dr_{24}}{2r_{24}}$$

Squaring, summing, dividing by N, we have

$$\frac{\sigma^2 r_{\infty\infty}}{r^2{}_{\infty\infty}} = \frac{\sigma^2 r_{12}}{r^2{}_{12}} + \frac{\sigma^2 r_{13}}{4r^2{}_{13}} + \frac{\sigma^2 r_{24}}{4r^2{}_{24}} - \frac{r_{12}r_{13}\sigma r_{12}\sigma r_{13}}{r_{12}r_{13}} - \frac{r_{12}r_{24}\sigma r_{12}\sigma r_{24}}{r_{12}r_{24}} + \frac{r_{13}r_{24}\sigma r_{13}\sigma r_{24}}{2r_{13}r_{24}}$$

We may obtain $r_{r_{12}r_{13}}$ and $r_{r_{12}r_{24}}$ by formula [129], $r_{r_{13}r_{24}}$ by formula [128], and all of the σ_r's by formula [108 *b*]. Doing so, collecting terms and simplifying yields,

$$\sigma r_{\infty\infty} = \frac{r_{\infty\infty}}{\sqrt{N}} \left\{ \frac{k^4{}_{12}}{r^2{}_{12}} + \frac{k^4{}_{13}}{4r^2{}_{13}} + \frac{k^4{}_{24}}{4r^2{}_{24}} - \frac{k^2{}_{13}}{r_{13}}\left(1 - \frac{r_{13}}{2} - \frac{r^2{}_{12}}{1+r_{13}}\right) - \frac{k^2{}_{24}}{r_{24}}\left(1 - \frac{r_{24}}{2} - \frac{r^2{}_{12}}{1+r_{24}}\right) + \frac{r^2{}_{12}(1-r_{13})(1-r_{24})}{r_{13}r_{24}} \right\}^{\frac{1}{2}}$$

In the notation of this chapter this is

$$\sigma r_{\infty\infty} = \frac{r_{\infty\infty}}{\sqrt{N}} \left\{ r^2{}_{\infty\infty} + \frac{1}{r^2{}_{12}} + \left(\frac{1}{4r^2{}_{13}} - \frac{r^2{}_{13}}{4} + r_{13} - \frac{1}{r_{13}} - 1\right) + \left(\frac{1}{4r^2{}_{24}} - \frac{r^2{}_{24}}{4} + r_{24} - \frac{1}{r_{24}} - 1\right) \right\}^{\frac{1}{2}}$$

(Standard error of a coefficient of correlation calculated by formula 155 *a*) ..... [161]

If we let A_{1I} stand for the first parentheses and A_{2II} for the second we have

$$\sigma r_{\infty\infty} = \frac{r_{\infty\infty}}{\sqrt{N}} \left(r_{\infty\infty}{}^2 + \frac{1}{r^2{}_{12}} + A_{1I} + A_{2II} \right)^{\frac{1}{2}}$$

The quantities $1/r^2$ and A are tabled for different values of r, in Table XL.

When the corrected coefficient of correlation is calculated by formula [161 *c*], or by

$$r_{\infty\infty} = \frac{r}{\sqrt{r_{13}}\sqrt{r_{24}}} \quad \text{.....[161 } a\text{]}$$

in which $r = (r_{12} + r_{14} + r_{32} + r_{34})/4$, the standard error of $r_{\infty\infty}$ is smaller than given by [161]. Before calculating this standard error let us note that r may be expeditiously obtained by calculating the correlation between the sum of the two

tests in the first trait and the sum of the two in the second trait. We have:

$$r_{(1+3)(2+4)} = \frac{\Sigma (x_1 + x_3)(x_2 + x_4)}{\sqrt{\Sigma (x_1 + x_3)^2}\sqrt{\Sigma (x_2 + x_4)^2}} = \frac{r_{12} + r_{14} + r_{32} + r_{34}}{4\sqrt{(1 + r_{13})/2}\sqrt{(1 + r_{24})/2}}$$

$$= \frac{r}{\sqrt{(1 + r_{13})/2}\sqrt{(1 + r_{24})/2}}$$

so that

$$r = r_{(1+3)(2+4)}\sqrt{(1 + r_{13})/2}\sqrt{(1 + r_{24})/2} \quad \ldots\ldots [161\ b]$$

Thus r may be easily obtained from a knowledge of the reliability coefficients and of the correlation between the two sums. Assuming that the arithmetic average is as reliable as the geometric average, then $r_{\infty\infty}$ calculated by [161 *a*] has the same reliability as $r_{\infty\infty}$ obtained from

$$r_{\infty\infty} = \frac{(r_{12}r_{14}r_{32}r_{34})^{\frac{1}{2}}}{\sqrt{r_{13}^2}\sqrt{r_{24}^2}}$$ (Yule's form of Spearman's formula for correction for attenuation).. [161 *c*]

The standard error of $r_{\infty\infty}$ calculated by this formula may be obtained in a manner very similar to that given in [161]. It is, however, a lengthy procedure and will not be recorded here. In brief it involves taking logarithmic differentials, squaring, summing, dividing by N, substituting values as given by formulas [108 *b*], [128] and [129], collecting terms after assuming that $r_{12} = r_{14} = r_{32} = r_{34} = r$. The answer is

$$\sigma_{r_{\infty\infty}} = \frac{r_{\infty\infty}}{2\sqrt{N}}\left(4r^2_{\infty\infty} + \frac{1}{r^2_{\infty\infty}} + \frac{1 + r_{13} + r_{24}}{r^2} + \frac{1}{r^2_{13}} + \frac{1}{r^2_{24}} - \frac{4}{r_{13}} - \frac{4}{r_{24}} - 2\right)^{\frac{1}{2}}$$

(Standard error of a coefficient of correlation calculated by formula 161 *a* or formula 161 *c*).... [161 *d*]

Magnitudes $1/r^2$ are given in Table XL. Study of this formula shows that the error in the corrected coefficient is very frequently not at all large, being in fact much smaller than given by Spearman (1910). The disagreement in derivation above [161 *d*] and that given by Spearman (1910, equation 24, p. 294), lies in the fact that Spearman, following Filon, to whom part of the derivation is credited, used formula [128] throughout, whereas formula [129] should at times have been used. The realization that this standard error is smaller than previously

recognized should throw much new light upon the question of the specific or general nature of intellectual functions.

TABLE XL

r	$1/r^2$	A	r	$1/r^2$	A	r	$1/r^2$	A
.01	10000.	2389.	.36	7.716	− 1.521	.71	1.984	− 1.329
.02	2500.	574.	.37	7.305	− 1.541	.72	1.929	− 1.316
.03	1111.	243.	.38	6.925	− 1.556	.73	1.877	− 1.304
.04	625.	130.	.39	6.575	− 1.568	.74	1.826	− 1.291
.05	400.	79.	.40	6.250	− 1.578	.75	1.778	− 1.280
.06	277.78	51.84	.41	5.949	− 1.584	.76	1.731	− 1.267
.07	204.08	35.80	.42	5.669	− 1.588	.77	1.687	− 1.255
.08	156.25	25.64	.43	5.408	− 1.590	.78	1.644	− 1.243
.09	123.46	18.84	.44	5.165	− 1.590	.79	1.602	− 1.231
.10	100.00	14.10	.45	4.938	− 1.588	.80	1.563	− 1.219
.11	82.645	10.68	.46	4.726	− 1.585	.81	1.524	− 1.208
.12	69.444	8.14	.47	4.527	− 1.581	.82	1.487	− 1.196
.13	59.172	6.23	.48	4.340	− 1.576	.83	1.452	− 1.184
.14	51.020	4.75	.49	4.165	− 1.570	.84	1.417	− 1.173
.15	44.444	3.59	.50	4.000	− 1.563	.85	1.384	− 1.161
.16	39.062	2.669	.51	3.845	− 1.555	.86	1.352	− 1.150
.17	34.602	1.931	.52	3.698	− 1.546	.87	1.321	− 1.138
.18	30.864	1.332	.53	3.560	− 1.537	.88	1.291	− 1.127
.19	27.701	.843	.54	3.429	− 1.527	.89	1.262	− 1.116
.20	25.000	.440	.55	3.306	− 1.517	.90	1.235	− 1.105
.21	22.676	.106	.56	3.189	− 1.507	.91	1.208	− 1.094
.22	20.661	− .172	.57	3.078	− 1.496	.92	1.181	− 1.083
.23	18.904	− .405	.58	2.973	− 1.485	.93	1.156	− 1.072
.24	17.368	− .601	.59	2.873	− 1.474	.94	1.132	− 1.062
.25	16.000	− .766	.60	2.778	− 1.462	.95	1.108	− 1.051
.26	14.793	− .905	.61	2.687	− 1.451	.96	1.085	− 1.041
.27	13.717	− 1.023	.62	2.601	− 1.439	.97	1.063	− 1.030
.28	12.755	− 1.122	.63	2.520	− 1.427	.98	1.041	− 1.020
.29	11.891	− 1.207	.64	2.441	− 1.415	.99	1.020	− 1.010
.30	11.111	− 1.278	.65	2.367	− 1.402	1.00	1.000	− 1.000
.31	10.406	− 1.338	.66	2.296	− 1.390			
.32	9.766	− 1.389	.67	2.228	− 1.378			
.33	9.183	− 1.432	.68	2.163	− 1.366			
.34	8.651	− 1.467	.69	2.100	− 1.353			
.35	8.163	− 1.497	.70	2.041	− 1.341			

With probable errors available there is no excuse for the indiscriminate averaging of corrected coefficients having values above and below 1.00, yielding possibly an average nearly equal to one. If we have a corrected coefficient equal to .90

with probable error of .02, and a second equal to 1.10 with a probable error of .02, we may conclude that neither coefficient is a chance variation from 1.00, and further that the fundamental hypotheses of similar tests, lack of correlation between errors, etc., underlying the idea of a reliability coefficient, must be absent in the case of the data yielding the corrected coefficient 1.10. A corrected coefficient greater than 1.00 is just as absurd as a "raw" coefficient greater than 1.00, and if positively found, as for example, $1.10 \pm .02$, it demands a reëxamining of hypotheses as truly as would the latter were it found to be greater than 1.00. Only in case corrected coefficients differ from 1.00 by such small amounts that the value 1.00 is well within the likelihood of occurrence, judged by the probable errors of the corrected coefficients, is it sound to average several such corrected coefficients to secure a measure of general tendency?

Section 59. Estimates of True Scores and the Probable Errors of These Estimates

Formula [153 *a*] has value for very practical reasons. For example, suppose we know that the reliability of foremen's judgments of the expertness of mechanicians is .36, and suppose we have a trade test the score upon which correlates with the judgments of one foreman to the extent of .48, then, letting the foreman's judgments equal X_0 and the trade test score equal X_1 we have

$$r_{1\infty} = \frac{r_{10}}{\sqrt{r_{00}}} = \frac{.48}{\sqrt{.36}} = .80$$

Thus the correlation between a single test score and an average of the judgments of an infinite number of foremen would be .8. If the hiring of a mechanician is not so much for the purpose of satisfying a particular foreman as it is to secure expert workmen the correlation .80 is not only the one of theoretical importance, but is, in fact, the correct one to use in regression equations estimating expertness from trade test score. We would have, letting x_∞ = the foreman's true judgment of expertness and $\bar{x}_\infty$ the best estimate of it.

$$\bar{x}_\infty = r_{1\infty} \frac{\sigma_\infty}{\sigma_1} x_1 \qquad \text{(Regression of a true criterion upon a fallible score)} \ldots\ldots\ldots\ldots\ldots\ldots [162]$$

The correlation $r_{1\infty}$ is given above and σ_∞ is immediately available, for we have, letting s bscripts here indicate scores on successive comparable trade tests,

$$\sigma^2_a = \frac{\Sigma (x_1 + x_2 + \cdots x_a)^2}{N} = \sum_1^a \sigma^2_p + \sum_1^{a^2-a} r_{pq}\,\sigma_p\,\sigma_q \quad ..[163]$$

And if the σ's are equal and r stands for the average of all the inter-correlations between the tests this reduces to

$$\sigma_a = \sigma\sqrt{a + (a^2 - a)r} \quad \text{(Standard deviation of the sums of } a \text{ comparable tests)} .....[164]$$

or, dividing by a and now letting σ_a stand for the standard deviation of the average of a such tests we have,

$$\sigma_a = \sigma\sqrt{\frac{1 - r}{a} + r} \quad \text{(Standard deviation of the averages of } a \text{ comparable tests)} ..[165]$$

And finally if a approaches ∞

$$\sigma_\infty = \sigma\sqrt{r} \quad \text{(Standard deviation of the averages of an infinite number of comparable tests)} ..[166]$$

Since $\sigma_a < \sigma$, the standard deviation of the true ability of a group is less than the standard deviation of the group upon a single fallible measurement. Accordingly measures of dispersion based upon single tests are too great to represent the true distribution. Estimates of true dispersion are given by formula [166]. As is obvious from the derivation, σ and r in the right hand member should be determined from the same population, or at least from two populations which one would expect to be equally homogeneous. I have elsewhere (Kelley, 1919 meas.), used formula [166] in the process of obtaining a measure of true overlapping in ability of two groups.

Returning to formula [162] we obtain

$$\bar{x}_\infty = \frac{r_{10}\sigma_0\sqrt{r_{00}}}{\sqrt{r_{00}}\,\sigma_1} x_1 = r_{10}\frac{\sigma_0}{\sigma_1} x_1 \quad \text{(Regression of a true criterion upon a fallible score)} .....[162\ a]$$

The reader will of course notice that the right-hand member of this equation is the same as that of formula [91 *a*] which gives the regression of a fallible criterion upon a fallible score. We thus have,

$$\bar{x}_\infty = b_{01}x_1 \quad \text{(Regression of a true criterion upon a fallible score)} .....[162\ b]$$

and

$$\bar{x}_0 = b_{01}x_1 \qquad \text{(Regression of a fallible criterion upon a fallible score)} \ldots\ldots\ldots\ldots\ldots\ldots [91\ b]$$

or the estimated true score is the same as the estimated single score. This is, of course, as it should be, and further it leads to the interesting fact that the standard errors of estimate in the two cases are different. We have

$$\sigma_{0\cdot 1} = \sigma_0 k_{01} = \sigma_0 \sqrt{1 - r^2_{01}}$$

(Standard error of estimate of a fallible criterion by means of a fallible score) . . . . . [86]

$$\sigma_{\infty\cdot 1} = \sigma_\infty k_{\infty 1} = \sigma_0 \sqrt{r_{00}} \sqrt{1 - \frac{r^2_{01}}{r_{00}}} = \sigma_0 \sqrt{r_{00} - r^2_{01}}$$

(Standard error of estimate of a true criterion by means of a fallible score) . . [167]

Thus we are able to estimate the true criterion score with smaller error than the fallible criterion. This is very satisfying. It means that in general, trade tests, intelligence tests, etc., actually accomplish a more accurate classification of those examined than indicated by the correlation with the criterion, since the criteria used are regularly fallible. The reliability coefficient r_{00} is of necessity greater than r^2_{01}, but with excellent tests and poor criteria it may not be very much greater, so that errors of estimate in placement may be small, and in fact much smaller than usually conceived. As a practical consequence it is seen that a systematic error in a criterion is very vicious, but that the chance error has no consequence whatever except in the requiring of a larger population in order to establish results with equal certainty.

Section 60. Accuracy of Placement on Basis of a Single Score

If in formula [162 *a*] we make X_∞ the average of many such scores as X_1, we have

$$\bar{x}_\infty = r_{1I}\, x_1$$

or

$$\bar{X}_\infty = r_{1I}\, X_1 + (1 - r_{1I})\, M_1 \qquad \text{(Regression of a true score upon a fallible score of the same function)} . [168]$$

The reason the correlation coefficient has replaced the regression coefficient of equation [162 *b*] is because we are here dealing

with similar scores, implying equal standard deviations, so that $r_{1I} = b_{1I}$. The accuracy of this estimate of a true score is given by

$$\sigma_{\infty \cdot 1} = \sigma_1 \sqrt{r_{1I} - r^2_{1I}}$$

(Standard error of estimate of a true score by means of a single score of the same function).....[169]

This formula is very valuable as it enables a judgment as to the accuracy of placement. Let us be given an elementary school reading test, having a reliability coefficient of .8 and a standard deviation of 10 on test scores covering the *same range of talent as that from which the reliability coefficient was determined*. If the sixth grade norm, or average score, equals 30, the seventh grade norm 38, and the eighth grade norm 46, let us determine the standard error of placement of a pupil as classified on the basis of the test score. We will first estimate the pupil's true score by formula [168]. The standard error of the estimated true scores, $\overline{X}_{\infty}$, is given by formula [169].

$$\sigma_{\infty \cdot 1} = 10 \sqrt{.80 - .64} = 5.0 -$$

The standard error of placement of the child is 5 and the probable error of placement $3\frac{1}{3}$, or 42 per cent of the difference between grade means. The question raised and answered has not involved a criterion outside of the test itself. With reference to that capacity which is measured by the test, we can say that the error of classification is a certain percentage of the difference between norms; or, if the difference between grade norms is called a year's growth, a certain percentage of a year's growth. Much may thus be determined without a criterion and this procedure is generally to be preferred to dependence upon a criterion having a systematic error, such, for example, as would be the case were a teacher to systematically judge pulchritude, vivacity, or mere industry, as evidence of reading ability. In addition to the simplicity of the method just described it may be recommended from the standpoint of reliability. The standard deviation of estimated true scores (estimated by means of the regression equation) is $\sigma_{\infty \cdot 1}$, and the standard deviation of test scores is σ_1. Accordingly $\sigma_{\infty \cdot 1}/\sigma_1$ is a measure of the proportionate reduction of error in the

placement of an individual having a given test score, over random placement. The smaller this ratio the greater the reduction. This quantity has a very small probable error as will immediately be shown so that the proportionate improvement due to the use of a test can be very accurately determined.

Let $\sigma_{\infty \cdot 1}/\sigma_1 = i$ = the measure of improvement due to the use of the test. Noting that the correlation between r and r_2 equals 1.0 we have

$$i^2 = r - r^2 = r\,(1 - r)$$

taking logarithmic differentials,

$$\frac{2i\,di}{i^2} = \frac{dr}{r} - \frac{dr}{1 - r}$$

Squaring, summing, and dividing by N,

$$\frac{4\,\sigma^2_1}{i^2} = \frac{\sigma^2_r}{r^2} + \frac{\sigma^2_r}{(1 - r)^2} - \frac{2\,\sigma^2_r}{r\,(1 - r)}$$

$$\sigma^2_1 = \frac{\sigma^2_r\,(1 - 2r)^2}{4\,i^2}$$

$$\sigma_i = \frac{(1 + r)\,i\,|\,1 - 2r\,|}{2r\sqrt{N}}$$

(Standard error of the measure of improvement, over random classification, resulting from the use of a score of reliability $r\,[= r_{11}]$).....[170]

Note that if $r_{11} = .5$ this standard error becomes zero. In the derivation of the formula second and higher powers of errors have as usual been discarded. Their inclusion would show that the standard error of this ratio is a trifle above zero when $r_{11} = .5$. If the error in r_{11} is of the order .02 the square is .0004, which is the order of the discarded portion, so that no material error is introduced in the formula by the omission of second and higher powers of the errors in r_{11} if N is greater than 25. In fact, for ordinary values of r_{11} we have a remarkably small σ_i. We need not hesitate to place confidence in an obtained value of i, even though the probable error of the obtained r_{11} is rather disconcertingly large.

Section 61. Average Intercorrelation

The correlation r_{1I} has occurred in several of the preceding formulas. If but two series of comparable scores are available this correlation may be calculated in but one way, but if there are several comparable series or forms of a test, which have been given, there are many ways of calculating the reliability coefficient. Having five comparable series of measures x_1, x_2, x_3, x_4, x_5 there are 10 possible pairings of series from which to calculate a reliability coefficient. This would in itself be a rather laborious task, but if the standard deviations of the several series are equal, or approximately so, the average of these 10 correlations may be calculated in a single operation since formula [163] may be solved for r, giving

$$r_{1I} = \frac{\frac{\sigma^2_a}{\sigma^2} - a}{a^2 - a}$$

(Average intercorrelation between a series, whose means, and standard deviations, are equal). [171]

The magnitude a is the number of series combined, so that it only remains to calculate σ_a and σ. If scores for each individual on the a forms are added, a series of N scores is obtained whose standard deviation is σ_a. Further the (aN) separate scores may all be entered into a single distribution and the standard deviation, σ, calculated. Thus whenever the means and the standard deviations of several series are equal, it is practically as simple to calculate the average intercorrelation as to determine a single correlation. It will now be shown that when ranks instead of scores are involved the calculation of the intercorrelation is still more simple. We need σ^2_a and σ. It has already been determined in Section 53 that if there are N ranks, 1, 2, 3, . . . N, their mean equals $(N + 1)/2$ and their standard deviation

$$\sqrt{\frac{N^2 - 1}{12}}$$

Accordingly

$$\sigma^2 = \frac{N^2 - 1}{12}$$

Let S equal the sum of the a ranks for a given individual, then

$$\frac{\Sigma S}{N} = a\left(\frac{N+1}{2}\right)$$

and

$$\sigma^2{}_a = \frac{\Sigma S^2}{N} - \left[a\frac{(N+1)}{2}\right]^2$$

Substituting the obtained values for σ^2 and σ^2_a and simplifying gives

$$r_{1I} = 1 - \frac{a\,(4\,N + 2)}{(a - 1)\,(N - 1)} + \frac{12\,\Sigma\,S^2}{a\,(a - 1)\,N\,(N^2 - 1)},$$

(Average intercorrelation between a series of N ranks). . [172]

This formula may be illustrated by a problem drawn from the writer's material. Six judges, K, T, U, B, L, H, rank according to merit 12 answers to a given problem as follows:

Ranks Given by Judges

Answers	K	T	U	B	L	H	S	S^2
A	1	5	7	10	2	5	30	900
B	2.5	6	4	6	3	9	30.5	930.25
C	2.5	3	1	4	1	2	13.5	182.25
D	4	2	2	11	8	3	30	900
E	5	12	3	1	4	10	35	1,225
F	6	1	8	2	5	1	23	529
G	7	11	10	8	12	4	52	2,704
H	8	9	5	7	6	11	46	2,116
I	9	4	9	12	7	6	47	2,209
J	10	7	11	5	9	8	50	2,500
K	11	10	12	9	10	12	64	4,096
L	12	8	6	3	11	7	47	2,209
								20,500.50

$a = 6,\ N = 12,\ \Sigma\,S^2 = 20{,}500.50$

therefore, by formula [172], $r_{1I} = .3241$

Such a problem as finding the average intercorrelation between the ranks of English compositions when 50 compositions are ranked by 100 judges would require the calculation of 4950 correlation coefficients, if no short-cut were available. But by the method illustrated the work could be done after the tabulation sheet is available in the time that might be required for four or five coefficients of correlation.

Suppose for the data just given it is desired to find out who is the best judge. The data are, of course, too scant to answer the question but they will illustrate the method. We might find correlations r_{KS}, r_{TS}, r_{US}, etc., and consider that judge the best who agrees most closely with the composite ranking. These correlations would enable a ranking of the judges, but they would be spuriously high because the rank of the judge

himself is included in the S composite. We therefore desire either $r_{K(S-K)}$ the correlation of the judge with the composite, omitting himself, or $(r_{KT}+r_{KU}+\cdots r_{KH})/5$ the average of all the correlations of each judge with the others. If judgments are expressed in the form of rankings, standard deviations are equal. The formula derived below will apply not only when ranks are used, but to any case in which standard deviations are equal. Let σ = the common standard deviation of the rankings. Let r_{1S} represent the correlation between the ranking of one judge and the sum of the rankings of all the judges, including himself. Let $r_{1\,(S-1)}$ be the correlation between the ranking of one judge and the sum of all the rankings of the other judges. Let

$$\bar{r}_{1p}=\frac{r_{12}+r_{13}+\cdots+r_{1a}}{(a-1)}$$

represent the average correlation between rankings of judge (1) and the other judges, and let $\bar{r}_{pq}$ equal the average of all the intercorrelations between the ranks of the judges. Then

$$\bar{r}_{1p}=\frac{1}{a-1}\overset{a-1}{\underset{1}{S}}r_{1p}$$

where p takes all values from 1 to a except the value 1.

$$\bar{r}_{pq}=\frac{1}{a^2-a}\overset{a^2-a}{\underset{1}{S}}r_{pq}$$

where p takes all values from 1 to a, and q takes all values except the value p.

$$r_{1S}=\frac{\Sigma x_1(x_1+x_2+\cdots+x_a)}{N\sigma\sigma_a}=\frac{\sigma^2+(a-1)\bar{r}_{1p}\sigma^2}{\sigma\sqrt{a\sigma^2+(a^2-a)\bar{r}_{pq}\sigma^2}}$$

$$r_{1S}=\frac{1+(a-1)\bar{r}_{1p}}{\sqrt{a+(a^2-a)\bar{r}_{pq}}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots [173]$$

Solving for $\bar{r}_{1p}$ we have

$$\bar{r}_{1p}=\frac{r_{1S}\sqrt{a+(a^2-a)\bar{r}_{pq}}-1}{a-1}$$

(Mean correlation between one series and $(a-1)$ others, in case standard deviations are equal) . . . [174]

The requirement that means shall be equal is necessary in case formula [171] is used for the calculation of $\bar{r}_{pq}$. The notation

r_{1I} was used upon the assumption that the several series were similar, but note that r_{1I} of formula [171] and $\bar{r}_{pq}$ in formula [174] are identical in derivation. The average intercorrelation $\bar{r}_{pq}$ is to be calculated once for all by formula [171] or [172] and r_{1S} calculated by the ordinary formulas [90], [93], [94], [95] or [142] for each successive series.

$$r_{1(S-1)} = \frac{Sx_1\,(x_1 + x_2 + \cdots x_a - x_1)}{N\sigma^2\sqrt{a + (a^2 - a)\,\bar{r}_{pq} - 1 - 2\,(a-1)\,\bar{r}_{1p}}}$$

$$r_{1(S-1)} = \frac{(a-1)\,\bar{r}_{1p}}{\sqrt{(a-1) + (a^2 - a)\,\bar{r}_{pq} - 2\,(a-1)\,\bar{r}_{1p}}}$$

(Correlation between one series and the composite of $(a-1)$ others in case standard deviations are equal) ..... [175]

Formula [175] involves $\bar{r}_{1p}$ which is already given by formula [174]. Substituting we obtain

$$r_{1(S-1)} = \frac{-\,(1 - r_{1S}\sqrt{a + (a^2 - a)\,\bar{r}_{pq}})}{\sqrt{-1 + a + (a^2 - a)\,\bar{r}_{pq} + 2\,(1 - r_{1S}\sqrt{a + (a^2 - a)\,\bar{r}_{pq}})}}$$

(Correlation between one series and a sum or average of $(a-1)$ others if standard deviations are equal) ..... [176]

To illustrate these formulas we may study the rankings of the six judges K, T, U, B, L, H to answer the question; which judge agrees most closely with the composite rankings of the others: We have

$$\sigma_K = \sigma_T = \cdots = \sqrt{\frac{144 - 1}{12}} = 3.4521$$

$$\sigma_S = \sqrt{\frac{20500.5}{12} - \left[\frac{6\,(12+1)}{2}\right]^2} = 13.6885$$

$$\Sigma\, x_K S = (30 + 76.25 + \cdots) - 12\left[\frac{12+1}{2}\right]\left[\frac{6\,(12+1)}{2}\right]$$

$$= 454.00$$

$$r_{KS} = \frac{\Sigma\, x_K S}{N\sigma_K\sigma_S} = .8006$$

A similar determination of the other correlations gives the table

$r_{KS} = .8006$	$r_{BS} = .3086$
$r_{TS} = .6604$	$r_{LS} = .8006$
$r_{US} = .7504$	$r_{HS} = .6437$

These coefficients establish the order of agreement of each judge with the others, but they are spuriously high in that S includes the record of each judge himself. We will, therefore, knowing by previous calculating, $\bar{r}_{pq} = .3241$, use formula [176] to calculate $r_{K(S-K)}$ and other similar coefficients. We obtain

$$r_{K(S-K)} = .6752 \qquad r_{B(S-B)} = .0592$$
$$r_{T(S-T)} = .4777 \qquad r_{L(S-L)} = .6752$$
$$r_{U(S-U)} = .6019 \qquad r_{H(S-H)} = .4554$$

These correlations may be taken at their face value. It is seen that judges K and L agree most highly with the other judges, while judge B agrees scarcely at all with the average opinion of the others.

Section 62. The Effect of Different Ranges upon Correlation of Similar Measures

I have elsewhere pointed out (Kelley, 1921 rel.) that a coefficient of correlation should be interpreted in the light of the ranges of the traits measured. This is true of all correlations, but it may be most readily proven when dealing with reliability coefficients. To quote from the reference cited, making such slight changes as are necessary to conform to the present notation:

"The reliability coefficient is, however, not an entirely satisfactory measure of reliability, for it is affected by the distribution, in the trait measured, of the particular group studied. To secure a reliability coefficient of .40 from a group composed of children in a single grade is probably indicative of greater, not less, reliability than to secure a reliability coefficient of .90 from a group composed of children from the second to twelfth grades. If it is reasonable to assume that in terms of true ability the spread of talent is four times as great in the eleven grades as in a single grade, the correlation in the second case would need to be .914 in order to indicate as close a relationship as that shown by a reliability coefficient of .40 in the single grade. The following formula gives the relationship:

$$\frac{\sigma_\infty}{\Sigma_\infty} = \frac{\sqrt{r_{11}\,(1 - R_{11})}}{\sqrt{R_{11}\,(1 - r_{11})}}$$

(Relation between ranges in true ability and reliability coefficients). [177]

σ_∞ and Σ_∞ are the standard deviations of the two groups in terms of true ability, and r_{1I} and R_{1I} are the reliability coefficients of the two groups. Solving this equation for the case in which $\Sigma_\infty = 4\,\sigma_\infty$, and $r_{1I} = .40$, gives $R_{1I} = .914$

"If the standard deviations of scores in two groups are known, it is not necessary to make any assumption; for then the following formula applies:

$$\frac{\sigma}{\Sigma} = \frac{\sqrt{1 - R_{1I}}}{\sqrt{1 - r_{1I}}}$$ (Relation between ranges in obtained scores and reliability coefficients)..[178]

In this formula σ and Σ are the standard deviations of the scores in the two groups and r_{1I} and R_{1I} the reliability coefficients respectively. In passing, it may be noted that this equation is an excellent criterion for determining whether a test is equally effective in a range Σ as in another range σ; for, if the relationship just given does not hold within the probable error of the determination, it is evidence that higher correlation is found in one part of the range than in another."

The proof of the above formulas is simple. Let $\sigma_{1\cdot\infty}$ = the standard deviation of an array of single test scores corresponding to a given true score for the one range of talent and $\Sigma_{1\cdot\infty}$ the standard deviation for the second range of talent. By formula [86]

$$\sigma_{1\cdot\infty} = \sigma_1\sqrt{1 - r^2_{1\infty}}$$

but by formula [160], $r^2_{1\infty} = r_{1I}$ so that,

$$\sigma_{1\cdot\infty} = \sigma_1\sqrt{1 - r_{1I}}$$

Similarly

$$\Sigma_{1\cdot\infty} = \Sigma_1\sqrt{1 - R_{1I}}$$

but if the test is equally as effective in one range as in the other the standard deviations of the divergences of the single scores from the true scores are equal, i.e.,

$$\sigma_{1\cdot\infty} = \Sigma_{1\cdot\infty}\,*$$

so that

$$\frac{\sigma_1}{\Sigma_1} = \frac{\sqrt{1 - R}}{\sqrt{1 - r}}$$ (Relation between standard deviations and reliability coefficients obtained from two different ranges when the measure is equally reliable throughout the two ranges).....[178]

* The validity of this equation is briefly discussed by Holzinger (1921).

Formula [165] enables us to express the same relationship dealing with true standard deviations instead of those obtained from single tests. Substituting for σ_1 and Σ_1, we have

$$\frac{\sigma_\infty}{\Sigma_\infty} = \sqrt{\frac{r\,(1-R)}{R\,(1-r)}} \quad \text{(Relation between true measures of dispersion and reliability coefficients obtained in two different ranges, when the measure is equally reliable throughout the two ranges).....[179]}$$

The fact that correlation changes with range makes comparison between reliability coefficients difficult. If one worker reports a test as having a reliability coefficient of .40 and a second reports a reliability coefficient of .90 for a test purporting to measure the same function we are not warranted in concluding without further data that the second test is the more reliable. For this reason the reporting of standard errors of estimate of true scores is to be recommended, for these will not change with the range if the test is equally effective throughout the range. Knowing the standard errors of estimate we would still be unable to compare two tests, if there is no equating of the units of the one test in terms of the units of the other. If the first worker reports a standard error of estimate for his test of 10 units, and the second a standard error of 2 units, and if some method of equating the scores (see Chapter VI) enables one to say that 6 units in the first test are equivalent in range covered to one unit in the second, then we can definitely say that the first test is the more reliable, for $10/6 < 2/1$. More extended discussion of this point is given in (Kelley, 1921 rel.).

Section 63. The Effect of Different Ranges upon Correlation of Different Measures

In case two different series of measures are correlated it is usually not known just what is the nature of the curtailment or extension of the ranges of the two series which has been brought about by some selective agency. In illustration; individuals of one race are probably less variable with reference to general intelligence and also less variable with reference to memory ability than humanity in general. But how much the decrease in variability is, or whether it is the same in the two functions

is not known. The correlation between general intelligence and memory ability determined from a random sampling of one range would probably be smaller than the same correlation calculated from humanity in general, but a priori considerations would give but a poor estimate of how great the difference is. In such a case and without additional data a correction of the correlation as found in the one range to enable a comparison with a similar correlation as found in the second range is impossible. If, however, the nature of the curtailment is known and is upon the basis of one trait only we may derive a formula enabling a comparison of correlation coefficients obtained from different ranges. Note that one trait is arbitrarily curtailed (or extended) and that the other is affected only in a consequential manner. Let x be the variable, the distribution of which is curtailed, and let y be the other variable. In the non-curtailed, scatter diagram let us suppose the y arrays are homoscedastic and show rectilinear regression. The dropping out of certain of these arrays, or of random parts of certain of them, will not change the slope of the regression line nor the homoscedasticity of the y-arrays, but it may be expected to change both the slope of the other regression line and the scedasticity of the x-arrays. Thus, designating the constants of the uncurtailed distribution by capital letters and of the curtailed by small letters, we have

$$\sigma_{2\cdot 1} = \Sigma_{2\cdot 1} \text{ and } b_{21} = B_{21} \quad \ldots\ldots\ldots [180] \text{ and } [181]$$

but

$$\sigma_{1\cdot 2} \neq \Sigma_{1\cdot 2} \qquad b_{12} \neq B_{12}$$

$$\sigma_1 \neq \Sigma_1 \qquad \sigma_2 \neq \Sigma_2$$

and

$$r_{12} \neq R_{12}$$

By formula [56] we have

$$\sigma_{2\cdot 1} = \Sigma_{2\cdot 1} = \sigma_2 \sqrt{1 - r^2_{12}} = \Sigma_2 \sqrt{1 - R^2_{12}}$$

or

$$\sigma_2 k_{12} = \Sigma_2 K_{12} \quad \text{(Relation between correlations and } y\text{-standard deviations when } x\text{-ranges have been changed).} \quad [182]$$

Note that formula [178] is but a special case of [182] for by letting the first variable be a true score and the second variable a score upon a single test of the same function, formula [182] becomes formula [178]. We may relate the y-standard devia-

tions to the x-standard deviations and obtain a relationship between the correlation and the standard deviation of the curtailed distribution. By formulas [87] and [180] we have

$$\Sigma^2_2 = \sigma^2_{2\cdot1} + \Sigma^2_{\bar{y}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots [183]$$

also

$$\bar{y} = r_{12}\frac{\sigma_2}{\sigma_1}x \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots [184]$$

Squaring, summing and dividing [184] by the population gives, for the uncurtailed distribution,

$$\Sigma^2_{\bar{y}} = r^2_{12}\sigma^2_2\frac{\Sigma^2_1}{\sigma^2_1}$$

Substituting in [183]

$$\Sigma^2_2 = \sigma^2_{2\cdot1} + r^2_{12}\sigma^2_2\left(\frac{\Sigma_1}{\sigma_1}\right)^2 = \sigma^2_2\left[(1 - r^2_{12}) + r^2_{12}\left(\frac{\Sigma_1}{\sigma_1}\right)^2\right] \quad \ldots.[185]$$

Substituting this value of Σ^2_2 in formula [182], dividing by σ_2 and solving for R_{12} yields

$$R_{12} = \frac{r_{12}\dfrac{\Sigma_1}{\sigma_1}}{\sqrt{1 - r^2_{12} + r^2_{12}(\Sigma_1/\sigma_1)^2}} \quad \ldots\ldots\ldots\ldots [186]$$

which is the result obtained by Pearson (1903, inf.).
This may be written in the form

$$\frac{R_{12}}{K_{12}\Sigma_1} = \frac{r_{12}}{k_{12}\sigma_1}$$

(Relation between correlations determined from ranges whose standard deviations in the case of the curtailed measure are in the ratio Σ_1/σ_1) $\ldots\ldots\ldots\ldots\ldots$ [187]

The only assumptions underlying this derivation have been rectilinearity and homoscedasticity in the curtailed trait. The standard error in R when thus determined is given in formula [300]. The accompanying table is presented to give a concrete idea of the differences in correlation that may be expected due to differences in range:

TABLE XLI

$\frac{\sigma_1}{\Sigma_1}$	If $r = .1$ Then $R =$	$r = .2$ $R =$	$r = .3$ $R =$	$r = .4$ $R =$	$r = .6$ $R =$	$r = .8$ $R =$	$r = .95$ $R =$
.75	.133	.263	.387	.503	.707	.872	.971
.50	.197	.378	.532	.658	.832	.936	.987
.25	.373	.632	.783	.868	.949	.983	.997
.10	.709	.898	.953	.975	.991	.997	.9995

A situation in which the ratio of the standard deviations may be determined is when the curtailed distribution is a part of a normal distribution. We have already noted [181], that

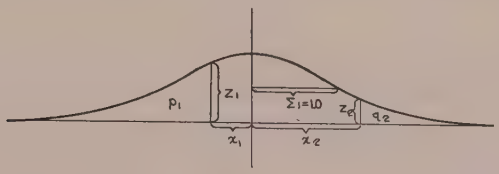
$$\frac{r_{12}\sigma_2}{\sigma_1} = \frac{R_{12}\,\Sigma_2}{\Sigma_1} = b_{21} = B_{21}$$

It is necessary to remember that the first variable x is the one upon the basis of which there has been a curtailment of distribution; that is, whatever difference there may be between σ_2 and Σ_2 is consequential to an imposed difference in σ_1 and Σ_1. This equation should be valuable in determining which of two functions is the more influential in causing selection. Suppose that for a narrow and a wide range we find b_{21} = approximately B_{21}, but that b_{12} does not = B_{12}. This suggests that trait (1) is the causal trait in bringing about the selection and trait (2) the consequential trait, or more accurately stated, that trait (1) is more closely related to whatever is the cause of the selection than is trait (2). Here again the regression coefficient is the significant constant for purposes of interpretation.

Brown (Brown, Carl — see Yerkes, 1921, pp. 629–632) has utilized certain properties of the normal distribution in determining the ratios of the standard deviations and therefore in determining the correlations in the two ranges. The Division of Psychology of the Surgeon General's Office found that many of its intelligence tests showed evidence of a "jam" at one or the other extreme; that is, the test was too difficult, resulting in large numbers of zero scores, or too easy, resulting in large numbers of perfect scores. Except for the extreme scores most of the tests gave approximately normal distributions. Accordingly the extremes of each test distribution were cut off and the correlation for the resulting scatter diagram calculated. This is an r from a curtailed distribution. If the ratio σ_1/Σ_1 can be determined, formula [186] will give the correlation R that would maintain throughout the entire distribution if the undistributed extreme scores could be replaced by scores as discriminative as those in the middle region of the distribution. We can obtain σ_1/Σ_1.

Let us be given a normal distribution of standard deviation

Σ_1 and cut off a proportion p_1 at the lower end and a proportion q_2 at the upper end, leaving a population of $(1 - p_1 - q_2)$, which is the same as $(q_1 - q_2)$ in the usual notation as given in Sections 24 and 27, from which the correlation r is obtained. No curtailment, except consequential, is made in variable 2. Let us suppose that the standard deviation of the non-truncated normal distribution Σ_1 is equal to 1.0. Then σ_1 as a proportion of Σ_1 is the only constant needed in order to use formula [186]. The standard deviation of that portion of the distribution, as shown in the accompanying diagram, lying



between the ordinates x_1 and x_2 is required. If the equation of the total normal distribution is

$$z = z_0 e^{\frac{-x^2}{2}}$$

the standard deviation of the truncated portion is given by

$$\sigma^2_1 = \frac{\int_{x_1}^{x_2} z\, x^2\, dx}{q_1 - q_2} - d^2$$

$$\int_{x_1}^{x_2} z\, x^2\, dx$$

integrated by parts and evaluated at the limits gives

$$x_1 z_1 - x_2 z_2 + (q_1 - q_2)$$

while d, the distance from the mean of the portion to the mean of the total normal distribution, is given by formula [55] so that

$$\frac{\sigma^2_1}{\Sigma^2_1} = 1 + \frac{x_1 z_1 - x_2 z_2}{q_1 - q_2} - \left[\frac{z_1 - z_2}{q_1 - q_2}\right]^2$$

(Standard deviation squared of a portion of a normal distribution of standard deviation, Σ_1, equal to 1.0) . . . . . [188]

Brown has called the right hand member $1 + J$ and introduced J into the equation giving r. We will, however, leave formula [186] as it is and expect σ_1/Σ_1 to be calculated by the present

formula [188] in case of truncation at one or both ends of a normal distribution and the resulting value introduced into formula [186]. Many very neat illustrations of the aid in interpretation resulting from the use of this formula are given in Yerkes (1921). One word of caution is offered. If multiple correlation coefficients are being calculated it is absolutely necessary that all the data be consistent. Otherwise such absurdities as imaginary correlation coefficients may result. Presumably if there are several variables, and every time a variable enters a correlation table its distribution is curtailed in one certain manner, not only would the r's, or the correlation from these truncated distributions be consistent with each other, but also the R's, or the enlarged correlations found by correcting for limited ranges. I have not proven this statement, but the converse is certainly obvious, that if the cut occurs in several places in the several scatter diagrams involving a certain variable there is no statistical imposition making the r's consistent, so that both the r's and the R's may be inconsistent. On page 633 of Yerkes (1921) occurs a table showing that army intelligence test Alpha_1 was cut between scores one and two in one scatter diagram and not cut at all in the other correlation tables. There is no evidence that for these particular data any inconsistency has been introduced by this procedure, but if the correlation had run high, .990–.999, instead of being less than .98 the lack of a necessary consistency in the original data would be serious.

Section 64. The Effect of Double Selection upon Correlation of Different Measures

A correction formula is available in case there has been selection in both variables. For example, consider a correlation between heights of brothers and sisters when brothers between heights a and b are used and when sisters between heights c and d, thus dropping out all pairings in which the brother's height lies outside of $\overline{ab}$, irrespective of sister's height, and also all pairings in which the sister's height lies outside of $\overline{cd}$ irrespective of the brother's height. Here there is selection both in the x trait and in the y trait. Let σ_1 and σ_2

be the standard deviations in the unselected distribution and let the selection in x alone be such as to change σ_1 to s_1, and let the selection in y alone be such as to change σ_2 to s_2. Let Σ_1 be the standard deviation of the x's and Σ_2 the standard deviation of the y's in the doubly selected distribution. To point the relation between s_1 and Σ_1 we may write

Σ_1 = the standard deviation consequent to the direct selection of the x's and also due to the indirect effect of selection of the y's.

s_1 = the standard deviation consequent to the direct selection of the x's.

Thus s_1 is not a standard deviation determined either from the original or the doubly selected population. It may, however, be determined by formula [188] or otherwise, if the nature of the selective agency operating upon the x's is known. The symbols s_2 and Σ_2 have similar meanings when dealing with the y's. Pearson (1908, inf.), starting with an original, normal correlation surface has given formulas showing the effect of double selection upon means, standard deviations, and correlation. Letting $t_1 = s_1/\sigma_1$, $t_2 = s_2/\sigma_2$ and letting small letters represent constants in the unselected distribution and capital letters in the selected, his formulas may be expressed:

$$\Sigma_1 = f(\sigma_1, t_1, t_2, r) \qquad \text{Given by Pearson (1908 inf.)} \quad [189]$$

$$\Sigma_2 = f(\sigma_2, t_2, t_1, r) \qquad \text{Given by Pearson (1908 inf.)} \quad [189\,a]$$

$$m_1 = \phi(\sigma_1, \sigma_2, m_1, m_2, t_1, t_2, r) \qquad \text{Given by Pearson (1908 inf.)} \quad [189\,b]$$

$$m_2 = \phi(\sigma_2, \sigma_1, m_2, m_1, t_2, t_1, r) \qquad \text{Given by Pearson (1908 inf.)} \quad [189\,c]$$

$$R_{12} = r_{12}\frac{t_1 t_2}{\sqrt{1 - r^2_{12}(1 - t^2_1)}\sqrt{1 - r^2_{12}(1 - t^2_2)}}$$

(Relation between r in a normal correlation surface and R in the surface obtained from the preceding by double selection) . . [190]

Theoretically one could solve equations [189] and [189 a] for t_1 and t_2 in terms of σ_1, σ_2, Σ_1, Σ_2 and r; substitute in formula [190] and thus relate R with r knowing the unselected and selected standard deviations. However, a solution of the t's in terms of the other constants runs into a bi-quadratic which apparently does not simplify so that the symbolic solution is not here attempted. The numerical solution for a given

problem is however possible, so that knowing Σ_1 and Σ_2 the ratios t_1 and t_2 may be determined from equations [189] and [189 *a*] or more simply, if the necessary facts as to curtailment are known, by formula [188], and substituted in formula [190] to obtain R.

Standard deviations may be either increased or decreased by selection due to increasing or decreasing certain arrays. Accordingly there is no necessity that t_1 or t_2 be less than one, nor that R be less than r. Whereas both the regression lines in the correlation surface or scatter diagram giving r are rectilinear since normality of surface was assumed, in general neither regression in the scatter diagram giving R will be rectilinear. As a consequence formula [190] is not symmetrical with reference to R and r. Selection could conceivably be of such sort that both the selected and unselected surfaces were normal, in which case the appropriate formula would of necessity be symmetrical with respect to R and r. The nature of the selection which would lead to this result is worthy of investigation.

CHAPTER X

FURTHER METHODS OF MEASURING RELATIONSHIP

Section 65. The Various Ways of Measuring Relationship

The treatment of the preceding two chapters has shown something of the extent and detail of analysis of inter-relationship between two quantitative variables which are related in a rectilinear manner, or at least in such a manner that a simple transformation will bring about rectilinear regression. If quantitative data are not of this nature, or if the data are qualitative, a number of accessory methods of measuring relationship are available, none of them, however, permitting the detail of interpretation and flexibility of treatment possible with rectilinearly related quantitative variables. Three general lines have been followed in developing accessory methods of measuring relationship: (1) leading to measures of relationship which would be identical with the product-moment correlation coefficient, provided data were (*a*) recorded in a quantitative instead of in a qualitative form and (*b*) related in a rectilinear instead of a curvilinear manner; (2) devising other measures of relationship; and (3) interpreting relationship in terms of probability.

The only method of the second and third groups which has, beyond cavil, demonstrated itself to be generally serviceable is the "goodness of fit" method developed by Pearson (1900, crit.). However, before treating of these methods we may concern ourselves with (1) the measures of relationship which are equivalent in meaning to the product-moment coefficient of correlation.

Section 66. The Median Ratio Correlation Coefficient

A method has been proposed by Thorndike (1913), which has not as yet been studied sufficiently to establish its compara-

bility with the product-moment coefficient for a variety of types of scatter diagrams. In the usual notation

$$r_{(\text{mdn ratio})} = \text{Median of } \left(\frac{x/\sigma_1}{y/\sigma_2}, \frac{y/\sigma_2}{x/\sigma_1}\right) \text{ ratios}$$

(Thorndike's median ratio coefficient of correlation) . . . . . [191]

In using this method some convention must be adopted with reference to x/o, y/o, and o/o ratios. In case grouping is fine, so that there is the possibility of few such ratios, the point is not important; but if there are large numbers of measures in the intervals having the means as their class indexes, then x/o, y/o and o/o combinations will make for uncertainty in results. Calling $1/2$ of these equal to ∞ and the other half equal to $-\infty$ will throw the burden of determining r upon the remaining ratios and, at least in the case of a normal correlation surface, this would not introduce a systematic error. If the grouping is fine so that the $x = o$ and $y = o$ frequencies are lacking or negligible in number, and if the correlation surface is normal, then the median ratio for any array is equal to the product-moment correlation coefficient, and, of course, the median of the ratios for the entire table equals the product-moment coefficient. We thus see that for this important correlation surface, and with fine grouping, Thorndike's median ratio coefficient has the same value as the product-moment coefficient. Further investigation of this coefficient is needed and, pending it, the method should not be used indiscriminately as a substitute for the product-moment method.

The distribution of ratios is very peculiar and the standard deviation of such distribution will generally be infinite, so that it is futile to calculate the standard error of the median ratio coefficient of correlation. The quartile deviation of these ratios, however, is not infinite, and we may take as a first approximation to the probable error,

$$\text{P. E. of } r_{\text{mdn ratio}} = \frac{\text{quartile deviation of ratios}}{\sqrt{N}}$$

(Approximate quartile error of the median ratio coefficient of correlation) . . . . . [192]

Noting that

$$\frac{x'/\sigma_1}{y'/\sigma_2} \cdot \frac{y''/\sigma_2}{x''/\sigma_1} = \frac{x'}{y'} \cdot \frac{y''}{x''}$$

the median of the

$$\left[\frac{x/\sigma_1}{y/\sigma_2}, \frac{y/\sigma_2}{x/\sigma_1}\right]$$

ratios will be closely equal to

$$\sqrt{(\text{mdn of } x/y \text{ ratios})\ (\text{mdn of } y/x \text{ ratios})}$$

Thus, we will write, as a very much simpler formula to use,

$$r_{\text{mdn ratio}} = \sqrt{(\text{mdn of } x/y \text{ ratios})\ (\text{mdn of } y/x \text{ ratios})}$$

(Thorndike's median ratio coefficient of correlation). . [191 *a*]

There is a certain directness in interpretation which commends this coefficient, but even in the form [191 *a*] it will hardly prove more expeditious to use than the regular product-moment method, while its probable error will, for usual surfaces, always be larger than the probable error of the product-moment coefficient.

Let us try this method upon the very curvilinear insurance data of Chart XXVII. We will use ξ and ζ as though they were x and y, deviations from the actual means, for comparison with our other calculations in which they were so used. We have the ratios listed below taking the measures by rows beginning at the top row. The calculation has been made by a slide rule, so that one need not expect an exact check upon every figure.

TABLE XLII

f	$\frac{\xi}{\zeta}$	$\frac{\zeta}{\xi}$	f	$\frac{\xi}{\zeta}$	$\frac{\zeta}{\xi}$		
1	12.9	.078	1	− 3.0	− .333		
1	12.2	.082	1	− 2.7	− .3750		
1	14.0	.071	2	− 2.4	− .417		
1	9.8	.102	1	− 2.2	− .458		
1	8.3	.120	1	1.6	.616		
2	7.5	.133	1	2.74	.365 +		
2	6.8	.147	2	− 17.3	− .058		
1	− 8.3	− .120	1	1.6	.628		
1	7.1	.140	2	2.2	.457 −		
3	5.0	.200	1	3.7	.271		
1	− 5.2	− .193	1	11.7	.086		
1	4.4	.226	1	3.0	.330		
1	3.9	.258	1	5.10	.196		
1	3.4	.290	1	16.2	.062		
2	2.8	.355	2	− 32.3	.031		
1	.3	3.000	1	2.61	.383		
2	− .25	− 4.000					
2	− .2	− 4.500					
2	− .2	− 5.500				Products	Square roots
1	− 8.0	− .125		L. Q. − 1.215	− .159	− .193 +	− .4395
				Mdn 2.675	.111	.297	.545 +
				U. Q. 5.95	.2645	1.573	1.255

$r_{\text{mdn ratio}} = .545$ Quartile error of $r = \frac{.847}{\sqrt{48}} = .122$

This result, $r = .55 \pm .12$, may be compared with the product-moment correlation, $r = .64 \pm .06$, and the corrected correlation ratios, $\eta_{12} = .73 \pm .05$ and $\eta_{21} = .74 \pm .05$. Thus for this particular surface in which the regression lines do not pass through the intersection of the means, the median ratio correlation is less than the product-moment correlation. Thorndike (1913) gives an illustration in which the median ratio coefficient is 1.00 and the product-moment coefficient less than 1.00. No general rule for the relation between these two correlations for non-rectilinear and non-homoclitic surfaces is offered.

Section 67. Correlation Determined from a Curve of Correspondence by Rank

This method, which may, more briefly, be described as the rank relation method, is proposed by Otis (1916). It probably has no essential advantages for rectilinear data, but offers promise if regressions are curvilinear. Having a scatter diagram, a line is to be drawn which will equate scores of the two variables. If regressions are rectilinear this line is given by the equation $x/\sigma_1 = y/\sigma_2$ (see Section 43), but if not rectilinear some other device must be followed. Otis writes (1916, p. 720): "In order to get a better idea where to draw the curve of relation an auxiliary plot may be made . . . on the assumption that the true correspondence of the scores of the two tests would be more nearly approximated by that of two scores having the same rank than by those of the same child." Otis does this graphically, smoothing slight irregularities. Having this curve of correspondence by rank we may locate a value on the x-scale for each value of y (or vice versa) and call the obtained value y'; that is, y' is, in terms of the x-scale, the equivalent of y. Thus y' measures and x measures have the same variability and the same mean. Let us designate the difference $(x - y')$ by the symbol d_x and designate $(y - x')$ by d_y. This enables us to use formula [131] in the calculation of the correlation. Otis notes that σ_{d_x}/σ_x is approximately equal to

$$\frac{\text{mdn of the } |d_x\text{'s}|}{\text{mdn deviation of } |x\text{'s}|}$$

so that, in our notation,

$$r = 1 - \frac{(\text{mdn of } |d_x\text{'s}|)^2}{2\,(\text{mdn dev. of } |x\text{'s}|)^2} \qquad \text{(Otis' deviation formula for correlation)} \ldots\ldots [193]$$

or, if x-values have been transformed into equivalent y-scores,

$$r = 1 - \frac{(\text{mdn of } |d_y\text{'s}|)^2}{2\,(\text{mdn dev. of } |y\text{'s}|)^2} \qquad \ldots\ldots [193]$$

These two formulas are minor modifications of formula [131], but Otis' manner of determining the d's is unique. These are not $(X - Y)$'s nor even $(x - y)$'s, but differences when (a) unequal variability has been allowed for, and (b) when one variable is transformed into a second by means of a curvilinear relation line. Thus the so-called r obtained is in reality more closely related to a correlation ratio η than to the correlation coefficient r, but it has an advantage over η, in that not only is the strength of the relationship measured, but the nature of it graphically established. The method suffers with all graphic methods in not enabling a concise algebraic statement of the relations which hold. We may expect the values obtained by its use to more nearly approach corrected η [200 b] than the product-moment r.

The insurance data of Chart XXVII may be used to illustrate the method, but to make it a little more algebraic than graphic we will equate measures by the method of Section 35, that is, we will call equal percentile values equivalent and will not resort to smoothing.

TABLE XLIII

Correspondence of Measures by Rank		Per Cent White Population	Per Cent White Population Rank Equivalent of Paired Insurance in Force Measure	Insurance in Force	Insurance in Force Rank Equivalent of Paired Per Cent White Population Measure
Insurance in Force	Per Cent White Population				
(a)	(b)	(c)	(d)	(e)	(f)
341	99	99	99	341	294
321	99	99	99	285	294
304	99	99	99	270	294
Mean 290	99	99	97	219	294
294 285	99	99	96	192	294
272	99	99	95.5	190	294
272	99	99	90	170	294
270	99	99	97	224	294

TABLE XLIII — *Continued*

Correspondence of Measures by Rank		Per Cent White Population	Per Cent White Population Rank Equivalent of Paired Insurance in Force Measure	Insurance in Force	Insurance in Force Rank Equivalent of Paired Per Cent White Population Measure
Insurance in Force	Per Cent White Population				
(*a*)	(*b*)	(*c*)	(*d*)	(*e*)	(*f*)
269	98	98	99	321	247
254	98	98	98	290	247
253	98	98	99	272	247
251	98	98	98	269	247
247	98	98	98	253	247
241	98	98	98	244	247
237	98	98	98	241	247
237	98	98	95	182	247
234	98	98	93	171	247
227	97	97	99	272	216
224	97	97	98	234	216
216	97	97	97	204	216
207	97	97	96	197	216
204	97	97	95	182	216
202	96	96	98	237	195
197	96	96	96	202	195
195	96	96	95·5	190	195
190	96	96	94	176	195
95·5					
190	95	95	99	304	185
185	95	95	98	251	185
182	95	95	98	237	185
176	94	94	82	140	176
171	93	93	56	103	171
170	90	90	88	167	170
167	88	88	83	142	167
158	87	87	57·5	105	158
147	84	84	98		
142	83	83	97	254	147
				207	142
140	82	82	97	227	140
136	82	82	54	101	133
133	80			133	133
133	78	78	80		
132	71	71	44	96	132
126	68	68	67	121	126
121	67	67	87	158	121
105	58	58	80	133	105
105	57·5	57	57·5	105	105
103	56	56	68		
101	54	54	84	126	103
96	44	44	71	147	101
84	43	43	43	132	96
				84	84

The measures in column (*a*) are insurance in force scores arranged according to magnitude, and the measures in column (*b*) per cent white population scores arranged according to magnitude. Column (*c*) is the same as column (*b*) and is obtained from the first column of Table XXXVII. The first entry, 99, in column (*d*) is the column (*b*) equivalent of 341 column (*a*), which is the measure paired with the first 99 in Table XXXVII. As a second illustration; the fifth 99, first column, Table XXXVII, is paired with 192. The value 192, column (*a*), is equivalent to 96, column (*b*), which is accordingly the value recorded in column (*d*) opposite the fifth 99 in column (*c*). The mean of column (*d*) is equal to that of column (*c*) and except for the slight grouping error in replacing 96 and 95 by 95.5 and 95.5, the replacing of 82 and 78 by 80 and 80, and the replacing of 58 and 57 by 57.5 and 57.5 the standard deviations are equal, so that we may use formula [131] in calculating the correlation. This gives $r = .70$.

A similar calculation, interchanging the variables, gives columns (*e*) and (*f*) and the final correlation $r = .65$. Compare this with $r = .64$, $\eta_{12} = .73$ and $\eta_{21} = .74$ of Section 52. These two correlation coefficients, or correlation ratios as they are more closely related to η than to r, should be differently labeled. Otis did not point out the fact that there are two for each table and that in general they will not be equal. The method is still in the elementary stage and needs (*a*) relating with r and with η, (*b*) an algebraic method (such as here used in equating percentiles, or still better a method resulting in the equation of the line of rank relation) for determining the curve of relation by rank, (*c*) determination of the types of correlation surfaces to which applicable, (*d*) utilization of coefficient and relation line obtained to estimate one variable knowing the second, and (*e*) determination of the probable errors of the constants involved. The most interesting feature of the method is that but a single relation line is used. However, the physical significance of this line will probably not be found to be as definite or serviceable as the regression lines of a correlation table.

Section 68. Correlation Ratio Method

Formula [86] gives the relationship between standard deviations of arrays and total standard deviation, and the coefficient of correlation in the case of rectilinear regression. Solving this for r^2 we have

$$r^2 = \frac{\sigma^2_1 - \sigma^2_{1\cdot 2}}{\sigma^2_1}$$

Formula [87] shows that, $\sigma^2_1 - \sigma^2_{1\cdot 2} = \sigma^2_{\bar{x}_y}$, leading to

$$r = \frac{\sigma_{\bar{x}_y}}{\sigma_1}$$

and also

$$r = \frac{\sigma_{\bar{y}_x}}{\sigma_2}$$

That is, if the regression is rectilinear the correlation coefficient is the ratio of the standard deviation of the means of the x-arrays to the standard deviation of the x's; or it is the ratio of the standard deviation of the means of the y-arrays to the standard deviation of the y's. This form suggests the use of these ratios when regressions are not rectilinear. The resulting values are called correlation ratios and are represented by the symbol η, eta, and note that there are two for each scatter diagram.

$$\eta_{12} = \frac{\sigma_{\bar{x}_y}}{\sigma_1} = \sqrt{1 - \frac{\sigma^2_{ax}}{\sigma^2_1}} \quad \text{(Correlation ratio of } x \text{ upon } y) \ldots\ldots [194]$$

$$\eta_{21} = \frac{\sigma_{\bar{y}_x}}{\sigma_2} = \sqrt{1 - \frac{\sigma^2_{ay}}{\sigma^2_2}} \quad \text{(Correlation ratio of } y \text{ upon } x) \ldots [194\,a]$$

The correlation ratio is of necessity greater than zero and less than one. The proof of this is left as an exercise. Further, σ_{ax} is the standard deviation of the x-arrays around their means, whereas $\sigma_{1\cdot 2}$ is the standard deviation of the x-arrays around the best fit straight line. The contribution of each array to $\sigma_{1\cdot 2}$ will be greater than the contribution to σ_{ax} in case the mean of the array is not exactly upon the regression line. Therefore $\sigma_{ax} \leqq \sigma_{1\cdot 2}$ and as a consequence $\eta > |r|$, and $\eta^2 > r^2$. The difference between η^2 and r^2 is ζ and is a measure of non-rectilinearity of regression. Therefore the test for linearity is

$$\zeta = \eta^2 - r^2 \quad \text{(Test for linearity of regression)} \ldots\ldots [195]$$

We need the standard error of this magnitude. Blakeman (1905) gives it as

$$\sigma_\zeta = \frac{2}{\sqrt{N}}\left(\zeta\left[(1-\eta^2)^2-(1-r^2)^2+1\right]\right)^{\frac{1}{2}} \quad \text{(Standard error of the test for linearity)} \ldots [196]$$

or approximately,

$$\sigma_\zeta = 2\sqrt{\frac{\bar{\zeta}}{N}} \ldots [197]$$

if η and r are not very different.

The calculation of $\sigma_{\bar{x}_y}$ offers no difficulties. The mean for each array is calçulated and the standard deviation of these found, taking each mean as many times as there are measures in the array. If the population is small the data should be grouped so that at least two measures are found in each array. The scatter diagram on page 241 shows the grouping that may be employed for the insurance data of Section 52. The class marks to the nearest $1.00 in the insurance in force data, and to the nearest 1 per cent in per cent of white population, are the means, not the mid-points of intervals, of the measures grouped. The origins are, to the nearest $1.00 and 1 per cent, the means of the total population. Neglecting the slight error due to not keeping fractional parts of the $1.00 or parts of 1 per cent gives the table and calculation on page 241.

The coarseness of grouping affects the size of η. With grouping so fine that but a single measure is found in any array, η would then = 1.0 and of course would have no real significance. In order to obtain a reasonable value for η grouping should be sufficiently coarse to result in a fairly regular, although not necessarily straight regression line. Pearson (1911 cor) has pointed out that the significance of η should be judged not by its difference from zero, but by its difference from the value that is the most probable in case of zero correlation between the two variables. Or in other words he has pointed out that a correction to the raw eta is necessary. Since the standard deviation of means of arrays are of necessity positive, this value for finite populations is as a matter of chance greater than zero, and if the population dealt with is small and the grouping fine it may be very much greater. The chances are, not only in the case of the zero relation, but

whatever the relation, that the obtained η is larger than it would be from an infinite population. Let η be the obtained correlation ratio ${}_f\eta$ the most probable ratio from an infinite population. And let κ equal the number of arrays; then, when the frequencies in the arrays do not differ in a very extreme manner from each other we have, as given by Pearson,

$$ {}_f\eta^2 = \frac{\eta^2 - \frac{(\kappa - 1)}{N}}{1 - \frac{(\kappa - 1)}{N}} \qquad \text{(Eta corrected for too fine a grouping)} \ldots [198] $$

Coarse grouping was resorted to in the calculation of η just given for the purpose of eliminating as much as possible of the error coming from too fine a grouping. But even so the correction is not negligible, since

$$ {}_f\eta^2_{12} = \frac{(.7955)^2 - \frac{9}{48}}{1 - \frac{8}{48}} = .5344, \text{ or } {}_f\eta_{12} = .7310 $$

$$ {}_f\eta^2_{21} = \frac{(.8019) - \frac{9}{48}}{1 - \frac{8}{48}}, \qquad \text{or } {}_f\eta_{21} = .7394 $$

The correlation ratio does not enable an estimation of one variable, knowing a second, as does the regression equation. Its value lies in giving a sort of upper limit to correlation. The use of some curvilinear regression line or transformation line, as in the case of the insurance and per cent white population data of Section 52, may lead to an actual means of estimating one variable knowing a second. The correlation ratio is also valuable as used with the data just mentioned in leading to ζ, and to the standard error of ζ, thus determining the likelihood of violation of data by the assumption of a rectilinear or other definite regression line. The standard error of η is usually taken as

$$ \sigma_\eta = \frac{1 - \eta^2}{\sqrt{N}} \qquad \text{(Standard error of the correlation ratio)} \ldots . [199] $$

but if η is large, due to too fine a grouping and small population, the standard error as given by this formula is too small and a corrective factor is necessary.

TABLE XLIV

X Insurance in force	Y Per Cent White Population											f	ξ	$\bar{\zeta}_\xi$	$f(\bar{\zeta}_\xi)^2$
	44	56	69	82	91	95	96	97	98	99					
341											1	1	142	$\frac{21}{2}$	220.5
321									1			1	122		
297						1			1			2	98	$\frac{17}{2}$	144.5
274								1	2	2		5	75	$\frac{51}{5}$	520.2
249				1		1			3			5	50	$\frac{31}{5}$	192.2
230				1		1	1	1		2		6	31	$\frac{40}{6}$	266.67
197				1			2	2		2		7	− 2	$\frac{-50}{7}$	357.14
175					1		1	1	2	1		6	−24	$\frac{-51}{6}$	433.5
147		1	1		2							4	−52	$\frac{-45}{4}$	506.25
129	1	2	1	1								5	−70	$\frac{-133}{5}$	3537.8
102		1	1	1	2							5	−97	$\frac{-95}{6}$	1504.17
84	1											1	−115		
												48		48	7682.93
f	2	4	3	5	5	3	4	5	9	8					160.061
ζ	−44	−32	−19	− 6	3	7	8	9	10	11					$\sigma_{\bar{\zeta}_\xi} = 12.6515$
$\bar{\xi}_\zeta$	$\frac{-185}{2}$	$\frac{-289}{4}$	$\frac{-219}{3}$	$\frac{-88}{5}$	$\frac{-322}{5}$	$\frac{179}{3}$	$\frac{3}{4}$	$\frac{78}{5}$	$\frac{472}{9}$	$\frac{326}{8}$					$\sigma_{\bar{\xi}_\zeta} = 51.2760$
$f(\bar{\xi}_\zeta)^2$	17112.5	20880.25	15987.	1548.8	20736.8	10680.3	2.25	1216.8	24753.8	13284.5		48	126203.0		
													2629.23		

The usual calculation gives $\sigma_1 = 64.4611$ and $\sigma_2 = 15.7778$. As calculated in the accompanying table $\sigma_{\bar{x}_y} = 51.2760$ and $\sigma_{\bar{y}_x} = 12.6515$, leading to $\eta_{12} = \frac{51.2760}{64.4611} = .7955$ and $\eta_{21} = \frac{12.6515}{15.7778} = .8019$.

The correction in η for too fine a grouping grows smaller as the number of categories decreases and this is as it should be, but an improved result is not obtained by a very coarse grouping, as then a correction for too coarse a grouping becomes important. This is based on formula [102] and is the same sort of a correction as given in formula [103] for a correlation coefficient, calculated from the means of the classes. Letting ${}_c\eta_{xy}$ be the value of η corrected for use of class means it may be readily shown, as has been done by Student (1913), that,

$$ {}_c\eta_{xy} = \frac{\eta_{xy}}{r_{y_\gamma}} \quad \text{(Correlation ratio corrected for coarse grouping)} \ldots\ldots [200] $$

and

$$ {}_c\eta_{yx} = \frac{\eta_{yx}}{r_{x_\chi}} \quad \ldots\ldots\ldots\ldots\ldots\ldots [200\,a] $$

To apply the correction we need to know r_{x_χ} and r_{y_γ}. The correlation between the class means and the deviates is $r_x = \sigma_x/\sigma_\chi$, and for the second variable $r_{y_\gamma} = \sigma_y/\sigma_\gamma$. The standard deviations σ_x and σ_y have already been determined in the calculation of η_{xy} and η_{yx} respectively. Were a normal distribution assumed σ_x/σ_χ could be determined as in the last chapter, but, though practically it might lead to good results, it is theoretically unsound for most distributions from which η is calculated. For the ungrouped data here given σ_χ may be determined from the raw data. Calculation without grouping from Table XXXVII gives $\sigma_\chi = 64.6746$ and $\sigma_\gamma = 15.8646$. Accordingly

$$ r_{x_\chi} = \frac{64.4611}{64.6746} = .99670 \text{ and } r_{y_\gamma} = \frac{15.7778}{15.8646} = .99453. $$

Thus for the corrected correlation ratio we have

$$ {}_c\eta_{xy} = \frac{\eta_{xy}}{r_{y_\gamma}} = \frac{.7310}{.99453} = .7350 $$

$$ {}_c\eta_{yx} = \frac{\eta_{yx}}{r_{x_\chi}} = \frac{.7394}{.99670} = .7418 $$

The values calculated as σ_χ and σ_γ have not been entirely freed from a grouping error, particularly σ_γ, since percentages recorded in the fundamental table are to the nearest 1 per cent only. To correct further it would be necessary to make some

assumption as to the form of distribution. Plainly the assumption of a normal distribution for the percentages of white population will not be sound. On the assumption that the distribution may be represented by a series of trapeziums of equal base, Student (1913) shows that the corrective factor is $\sqrt{1 + h^2/(12\,\sigma^2)}$ in which h is the unit of grouping and σ the standard deviation of χ in the case of $\eta_{y\chi}$ and γ in the case of $\eta_{x\gamma}$. Applying this further correction to $\eta_{x\gamma}$ we have

$$_{cc}\eta_{x\gamma} = .7350\sqrt{1 + \frac{1}{12 \times 15.8646}} = .7352$$

This correction is merely a re-application of the $r_{y\gamma}$ division and is warranted due to the fact that division by .99453, the $r_{y\gamma}$ obtained, allowed only for the grouping of several percentages and not for the error introduced by entering values in the original table to the nearest per cent only. For the data in hand the only correction really worth while was the first, formula [198], that for too fine grouping. The second, that for too coarse grouping, will amount to 1 per cent if $h = \sigma/2$, or in the case of a normal distribution if there are some 10 or 12 steps, or intervals. This result is obtained by solving the equation

$$\sqrt{1 + \frac{h^2}{12\,\sigma^2}} = 1.01$$

A correction for grouping by means of Sheppard's formula [68 *a*] applied to the standard deviation in the divisor of the formula giving the raw η, is appropriate, but no such correction for the standard deviation in the dividend is to be made for this is a standard deviation of means, or points, and should not be corrected by Sheppard's formula which applies to continuous variates.

As there are so many corrections which apply to η the following summary is given,

Let $\sigma_{\bar{x}_y}$ = the standard deviation of the means of the x-arrays.

Let σ_x = the standard deviation of the x's.

Then letting η_{xy} equal the raw correlation ratio of the x's upon the y's we have

$$\eta_{xy} = \frac{\sigma_{\bar{x}_y}}{\sigma_x} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots [194]$$

Letting ${}_s\eta_{xy}$ equal the value after applying Sheppard's correction for grouping of the x's we have, if h equals the number of units per group,

$$ {}_s\eta_{xy} = \frac{\sigma_{\bar{x}_y}}{\sigma_x\sqrt{1 - \frac{h^2}{12\,\sigma^2_x}}} = \frac{\sigma_{\bar{x}_y}}{\sigma_x}\sqrt{1 + \frac{h^2}{12\,\sigma^2_x}} = \eta_{xy}\left(1 + \frac{h^2}{24\,\sigma^2_x}\right) \quad \text{..[194 } b\text{]} $$

Letting κ stand for the number of y categories, N the total number of cases and ${}_{fs}\eta_{xy}$ the preceding value of η corrected for too fine grouping of the y's, we have,

$$ {}_{fs}\eta^2_{xy} = \frac{{}_s\eta^2_{xy} - (\kappa - 1)/N}{1 - (\kappa - 2)/N} \quad \text{.....[198 } a\text{]} $$

Letting $r_{y\gamma}$ equal σ_y/σ_γ, i.e., the correlation between the class means of the y's and the y variates back of the grouped data (note that σ_y is the standard deviation of the class means, but that σ_x above [194 b] is the standard deviation of class indexes), and letting ${}_{cfs}\eta_{xy}$ equal the preceding η corrected for too coarse a grouping of the y's we have

$$ {}_{cfs}\eta_{xy} = \frac{{}_{fs}\eta_{xy}}{r_{y\gamma}} \quad \text{.....[200 } b\text{]} $$

In the case of equal intervals in y which are not too large $\left(\text{say not} > \frac{\sigma}{2}\right)$, $\sigma^2_\gamma = \sigma^2_y\left(1 + \frac{h'^2}{12\,\sigma^2_y}\right)$ in which σ_y is as before the standard deviation of means of y classes and h' the number of units per group of y's, so that $1/r_{y\gamma}$ then equals

$$ \left(1 + \frac{h'^2}{24\,\sigma^2_y}\right) $$

and we have

$$ {}_{cfs}\eta_{xy} = {}_{fs}\eta_{xy}\left(1 + \frac{h'^2}{24\,\sigma^2_y}\right) \quad \text{.....[200 } c\text{]} $$

In [200 c] we may substitute the standard deviation of class indexes for σ_y, the standard deviation of class means, without appreciable error, but we cannot make this substitution in the general formula, $r_{y\gamma} = \sigma_y/\sigma_\gamma$ [102], which is the formula which must be used in case the grouping of y's is in very broad and unequal intervals, and especially if the classes are categories not related in a numerical manner.

These corrections to η_{xy} are not equally demanded in the case of any given data. Correction [198] is likely to be the most necessary. The finer the y grouping, that is, the larger

the number of y-categories and the smaller the total population the more important is this correction. Correction [194 b] is important if the x-grouping is coarse and correction [200] if the y-grouping is coarse. All of these observations apply to η_{xy} and of course similar statements will hold with reference to η_{yx}, if in the statements y and x are interchanged throughout.

The student should note that the value of η used in the calculation of ζ, the test for linearity, and in the calculation of the standard error of ζ, is the raw value and not the corrected value. Although the corrected value of η should not be used in these formulas [195], [196], [197] as it was not involved in the derivation of ζ, nevertheless the formula for ζ calculated from raw η may be expected to give a value which is materially too large, and a value for its standard error which is relatively too small, if grouping is fine and population small. Accordingly the ζ test for linearity is too rigorous if grouping is fine and population small.

Section 69. Method of Parabolic Regression

Many scatter diagrams are characterized by regular curvilinear regression lines. If a single positive or negative curvature is present the regression line may sometimes be closely represented by a parabola, $y = a + bx + cx^2$; and if the regression line shows a single inflection the cubic parabola,

$$y = a + bx + cx^2 + dx^3$$

may give a good fit. Pearson (1905) has developed the theory of parabolic regression and illustrated the procedure with certain data. It is too involved to give here, but must needs be resorted to if the specific nature of the curvilinear regression line and the numerical values of the constants involved constitute the crux of the problem.

Section 70. Bi-Serial r Method

In case one series consists of variates, or graduated measures, and the other is dichotomous we may determine the correlation that maintains if we assume that the trait represented by the dichotomic distribution is in reality a continuous trait, normal in distribution, for which we have only categorical information. Such a situation is well represented by the following, taken

from the army psychological test data (Yerkes, 1921, p. 748). We may proceed with the steps involved in obtaining the numerical value of bi-serial r and consider the general formula afterward.

TABLE XLV

Score in Army Alpha Intelligence Test	Number of Men Who Left School	
	Below the 9th Grade	Above the 8th Grade
205–212		1
200–204		3
195–199		14
190–194		17
185–189	1	49
180–184	2	54
175–179	8	78
170–174	12	126
165–169	18	149
160–164	15	200
155–159	20	244
150–154	45	305
145–149	58	352
140–144	74	338
135–139	101	407
130–134	145	507
125–129	190	528
120–124	216	530
115–119	317	643
110–114	393	674
105–109	507	682
100–104	582	691
95– 99	761	712
90– 94	908	725
85– 89	993	769
80– 84	1,181	693
75– 79	1,371	642
70– 74	1,604	648
65– 79	1,709	567
60– 64	1,962	581
55– 59	2,249	430
50– 54	2,272	346
45– 49	2,429	305
40– 44	2,455	229
35– 39	2,473	200
30– 34	2,490	154
25– 29	2,213	106
20– 24	1,835	60
15– 19	1,511	42
10– 14	545	13
5– 9	432	5
0– 4	183	3
	34,280	13,822
	13,822	
	48,102	

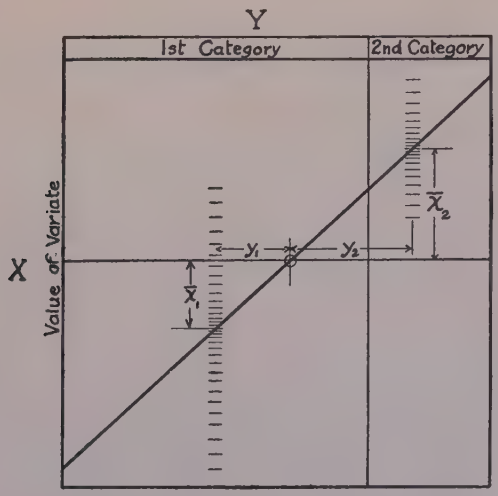
M_1 and M_2 are the means of the first and second categories respectively, and σ is the standard deviation of the total distribution (48,102) of measures. Calculation by methods already given yield

$$M_1 = 54.987, \quad M_2 = 98.758$$
$$\sigma = 36.606$$

and finally

$$r = \frac{M_2 - M_1}{\sigma} \times \frac{pq}{z} = .7435.$$

With this concrete calculation in mind we may turn to the more general statement of the problem. The army Alpha series is a variate series, and the graduation or non-graduation from the elementary school a categorical series, not corresponding to a true dichotomy in talent of any sort whatever. Even in terms of schooling the two classes are not homogeneous within themselves. In the non-graduation class are individuals who have been in school variously 0, 1, 2, . . . 8 years, while the completion of the elementary school class comprises those who have been in school 9, 10, . . . years. Thus the dichotomy has been arbitrarily imposed upon a continuous trait. Let X equal the scores in the variate trait and Y those in the dichotomous trait, then $r = b_{12}\frac{\sigma_2}{\sigma_1}$. The regression line



with slope b_{12} passes through the means of the x-arrays, $\bar{x}_1$, x_2, of the distribution of cases in the two y-categories. Therefore, referring to the diagram on page 247,

$$b_{12} = \frac{\bar{x}_2}{y_2} = \frac{\bar{x}_1}{y_1} = \frac{\bar{x}_2 + \bar{x}_1}{y_2 + y_1} \text{ and } r = \frac{(x_2 + x_1)}{\sigma_1} \Big/ \frac{(y_2 + y_1)}{\sigma_2}$$

Now $(\bar{x}_2 + \bar{x}_1)$ is simply $(M_2 - M_1)$ the difference between the means of the x-scores in the two categories, and σ_1, or simply σ, is the standard deviation of the total distribution of x-scores. It therefore only remains to obtain

$$\left(\frac{y_2}{\sigma_2} + \frac{y_1}{\sigma_2}\right)$$

Let p be the proportion of cases in the first y-category and q the proportion in the second. The distance y is simply the mean deviate of the tail of a normal distribution and is given by formula [83]. If z is the ordinate, as given in Table K–W, at the point of truncation of the normal distribution, cutting off p proportion of cases we have

$$\frac{y_1}{\sigma_1} = \frac{z}{p} \text{ and } \frac{y_2}{\sigma_2} = \frac{z}{q} \text{ so that } r = \frac{\dfrac{M_2 - M_1}{\sigma}}{\dfrac{z}{p} + \dfrac{z}{q}}$$

which may be written

$$r = \frac{(M_2 - M_1)\,pq}{\sigma z} \quad \text{(Bi-serial coefficient of correlation)} \ldots\ldots\ldots\ldots\ldots\ldots [201]$$

This formula differs somewhat from, and is more simple to use than Pearson's (1909), but is identical in the principle underlying its derivation. The coefficient as derived has been called "bi-serial r," and must be distinguished from "bi-serial η," described in the next section.

In case the grouping of x's is coarse, Sheppard's correction should be applied in determining σ. In case the population is small there is a chance correlation greater than or less than zero dependent upon the point of dichotomy, so that a correction of the value just given is necessary. Soper (1914 bi-ser) gives the following correction formula in which $_cr$ is the corrected value, r the value given by formula [201], x the

deviate given in Table K–W corresponding to area q, the proportion q being the smaller of the two proportions p and q.

$$r' = r\left\{1 + \frac{1}{N}\left[\frac{1}{4} + \frac{pq}{2z^2} - \left(1 - \frac{px}{z}\right)\left(1 + \frac{qx}{z}\right) + \frac{1}{2}r^2\right]\right\}$$

(Bi-serial r corrected for small population).....[202]

Note that for moderate dichotomies and populations greater than 100 this correction may generally be considered negligible. The square of the standard error of bi-serial r as given by Soper is

$$\sigma^2_r = \frac{1}{N}\left\{\frac{pq}{z^2} - \left[\frac{3}{2} + \left(1 - \frac{px}{z}\right)\left(1 + \frac{qx}{z}\right)\right]r^2 + r^4.\right\}$$

(Square of standard error of bi-serial r).....[203]

For dichotomies wherein q is not less than .05 a close approximation to the preceding formula is

$$\sigma_r = \frac{\left(\frac{\sqrt{pq}}{z} - r^2\right)}{\sqrt{N}}$$

(Standard error of bi-serial r).....[204]

Even for extreme dichotomies this last formula which gives a slightly larger value for σ_r than formula [203] may well be preferred, for the assumption of normality of distribution underlying formula [203] is generally less safe in the case of extreme than of moderate dichotomies, so that an increase in the size of the standard error due to the extra hazard of the assumption of normality is desired and this is given by formula [204]. Certain of the functions involved in formulas [202] and [203] have been tabled by Soper in the reference cited. The evaluation of these formulas is also readily accomplished by the aid of Table K–W.

Section 71. Bi-serial Eta

The title of the original contribution by Pearson (1910, new) describes the data to which this method applies: "On a new method of determining correlation where one variable is given by alternative and the other by multiple categories." To quote further from Pearson (1917 bi-ser.): "Let x be the alternative, y the multiple variate, $\bar{x}_y$ the distance from the division between the alternative categories of the mean of the array of x's corresponding to a given value of y, ${}_y\sigma_x$ its standard deviation and n_y its frequency. Let $\bar{x}$, σ_x and N be the cor-

responding quantities for the marginal totals." To utilize the notation of Table K–W, let

$$x = \frac{\bar{x}}{\sigma_x},\ x_y = \frac{\bar{x}_y}{y\sigma_x},\ \text{and}\ K^2 = \frac{1}{N} S(n_y x^2{}_y)$$

In the notation of Table K–W, x_y is the deviate corresponding to q_y, the proportion of cases lying above the point of dichotomy of the y-category, and x without subscript is simply the deviate corresponding to q, the proportion of cases constituting the smaller of the two x-categories. The number of cases in a y-category is n_y and S is a summation covering all the categories in the multiple category variate. Thus

$$\eta_{xy} = \left[\frac{K^2 - x^2}{1 + K^2}\right]^{\frac{1}{2}} \quad \text{(Bi-serial eta)} \ldots\ldots\ldots\ldots\ldots\ldots [205]$$

There is no correction to be made to this formula on account of the x-variate, but correction formula [198 *a*] should be used if κ, the number of y-categories, is large and the population, N, small; and correction [200 *b*] or [100 *c*] should be made if the number of y-categories is small. If η is small, so that higher powers are relatively unimportant with reference to η and η^2, the standard error of η is given by

$$\sigma_\eta = \frac{1 - \eta^2}{\sqrt{N}}\left(\frac{pq}{z^2} + \frac{2\,px^2}{(1 + x^2)^2}\right)^{\frac{1}{2}} \quad \text{(The standard error of a bi-serial } \eta \text{ which is equal to 0)} \ldots\ldots\ldots [206]$$

The magnitudes p, q, z, x are constants of the alternative category distribution having the usual meanings and are available from Table K–W when q is known. If η is greater than .5 the full formula for its standard error as given and fully described by Pearson (1917 bi-ser.), is needed.

We may use data comparing southern and northern negroes collected by the Division of Psychology of the Surgeon General's Office to illustrate the method. In general the army Alpha test was given to literate individuals of greater than feeble-minded intelligence, and army Beta or an individual test was given to illiterate individuals or to literate persons of very limited intelligence. Accordingly a division of individuals upon the basis of whether they were tested by means of army Alpha alone; or by means of army Alpha and Beta, or army Beta, or army individual, will constitute a dichotomy closely related to literacy. Table 4, pages 559–60 of Yerkes (1921), enables us to determine whether there is a correlation between

negro literacy and domicile as represented by State of the union. The table together with the supplementary columns used in the calculation of bi-serial η follows:

TABLE XLVI

Negro Draft — Pro-rated by States

State	Examination Taken		n_y	$q=\frac{n_\alpha}{n_\alpha+n_\beta}$	x_y	$n_y x^2_y$
	Alpha Only	Alpha-Beta, Beta, or Individual				
Alabama	271	1,088	1,359	.1994	.8452	970.82
Arizona	3	4	7	.4286	.1789	.22
Arkansas	192	706	898	.2136	.7926	564.14
California	31	28	59	.5254	−.0627	.23
Colorado	18	12	30	.6000	−.2533	1.92
Connecticut	17	28	45	.3778	.3107	4.34
Delaware	40	44	84	.4762	.0602	.30
Dist. of Col.	30	180	210	.1429	1.0669	239.04
Florida	499	122	621	.8035+	−.8560	455.03
Georgia	416	1,969	2,385	.1744	.9385	2,100.67
Idaho	4	8	12	.3333	.4316	2.24
Illinois	137	114	251	.5458	−.1156	3.35
Indiana	74	51	125	.5920	−.2327	6.77
Iowa	23	13	36	.6389	−.3558	4.56
Kansas	87	30	117	.7436	−.6557	50.30
Kentucky	191	341	532	.3652	.3451	63.36
Louisiana	538	1,147	1,685	.3193	.4705	373.01
Maine	0	0	0			
Maryland	146	379	525	.2781	.5888	182.01
Massachusetts	54	39	93	.5806	−.2045	3.89
Michigan	17	25	42	.4048	.2404	2.43
Minnesota	9	11	20	.4500	.1257	.32
Mississippi	773	967	1,740	.4443	.1408	34.49
Missouri	196	182	378	.5185+	−.0476	.86
Montana	2	2	4	.5000	.0000	.00
Nebraska	13	13	26	.5000	.0000	.00
Nevada	0	3*	3*			
New Hampshire	0	1*	1*			
New Jersey	105	72	177	.5932	−.2353	9.80
New Mexico	3	1	4	.7500	−.6745	1.82
New York	197	107	304	.6480	−.3799	43.87
North Carolina	211	1,168	1,379	.1530	1.0237	1,445.14
North Dakota	2	1	3	.6667	−.4316	.56
Ohio	163	88	251	.6494	−.3826	36.74
Oklahoma	98	211	309	.3172	.4761	70.04
Oregon	3	3	6	.5000	.0000	.00
Pennsylvania	183	236	419	.4368	.1586	10.54
Rhode Island	9	9	18	.5000	.0000	.00
South Carolina	334	1,303	1,637	.2040	.8274	1,120.68
South Dakota	1	15	16	.0625	1.5382	37.86
Tennessee	504	433	937	.5379	−.0954	8.53

* Omitted in totals.

TABLE XLVI — *Continued*

Negro Draft — Pro-rated by States — Continued

State	Examination Taken		n_y	$q = \frac{n_\alpha}{n_\alpha + n_\beta}$	x_y	$n_y x^2 y$
	Alpha Only	Alpha-Beta, Beta, or Individual				
Texas . . .	786	1,048	1,834	.4286	.1789	58.70
Utah . . .	4	5	9	.4545+	.1130	.11
Vermont . .	0	0	0			
Virginia . .	56	1,148	1,204	.0465+	1.6747	3,376.76
Washington .	7	9	16	.4375	.1560	.39
West Virginia .	67	101	168	.3988	.2559	11.00
Wisconsin . .	2	5	7	.2857	.5651	2.24
Wyoming . .	4	2	6	.6667	−.4316	1.12
	6,520	13,468	$19{,}988 = N$			11,300.20
				$q = .32620$, $x = .450431$		

$$z = .360457$$

$$\eta^2 = \frac{.362461}{1.565349}$$

$$\eta = .481199$$

$$\sigma_\eta = .006184$$

$$K^2 = .565349$$

$$x^2 = .202888$$

$$.362461$$

The bi-serial correlation ratio is less than .50 so that we may obtain a satisfactory idea of its probable error by using formula [206]. This gives a standard error of .00618 which is so small with reference to η as to establish the fact that there is a moderate correlation of about .48 between literacy of the negro and domicile. The obtained value should theoretically be corrected by applying formulas [198 *a*] and [200 *b*] or [200 *c*]. They are entirely inconsequential in this problem, but will be used to show the method. The number of categories in the y-variate is 45 (number of states yielding frequencies) so that we have, applying correction [198 *a*],

$$f\eta^2_{xy} = \frac{(.481199)^2 - 44/19988}{1 - 43/19988}$$

from which

$$f\eta_{xy} = .479423$$

This correction (.4812−.4794 = .0018) is not large, but even so it is probably somewhat too great as the 45 y-categories have

such extremely varying frequencies that the hypotheses underlying the correction are not closely met. The states constitute a geographical series and no assumption with reference to numerical relationship between them seems warranted, nor any assumption as to total distribution on a one dimensional scale. However, some correction for coarseness of grouping is appropriate. We will assume a rectangular distribution of states of equal populations and will not attempt to justify the assumption further than to say that the correction that it leads to is probably conservative, i.e., too small rather than too large, so that our procedure is an improvement over one not involving a correction. The standard deviation of a rectangular distribution of 45 ranks equals

$$\sqrt{(n^2 - 1)/12} = \sqrt{2024/12}$$

so that since the unit of grouping is the state, correction [200 *c*] is as follows: making $h' = 1.0$:

$$_{cf}\eta_{xy} = .479423\left(1 + \frac{1}{4048}\right) = .479541$$

The reader will understand that the number of figures to which the work has here been carried and the corrections made are for illustrative purposes only and that to meet practical demands the raw result, $\eta_{xy} = .481$, would be adequate for these particular data. We may now turn to a consideration of the correlation between two series, the measures of each of which lie in alternative categories.

Section 72. Tetrachoric Correlation

In case we have a 2×2 fold table such, for example, as is given by indicating the presence or absence of two traits we may calculate r_t, the tetrachoric coefficient of correlation. The assumption underlying the method is that both traits are really continuous and normal in distribution and that the dichotomies have forced the data for each trait into two alternative categories. The procedure was developed by Pearson (1900 cor.), and tables of "Tetrachoric Functions"

have been calculated by Everitt (1910 — also given in Pearson's Tables 1914 t). Pearson started with the 2 × 2 fold table,

TABLE XLVII

a	b	$a+b$
c	d	$c+d$
$a+c$	$b+d$	N

so arranged, as is obviously always possible, that $a+b>c+d$ and $a+c>b+d$. We will start with a table of the same sort dealing with proportions instead of gross numbers. Let

$$\alpha=\frac{a}{N},\ \beta=\frac{b}{N},\ \gamma=\frac{c}{N},\ \delta=\frac{d}{N}$$

$$p=\frac{a+b}{N},\ q=\frac{c+d}{N},\ p'=\frac{a+c}{N},\ q'=\frac{b+d}{N}$$

Then our table becomes

TABLE XLVIII

α	β	p
γ	δ	q
p'	q'	1.0

Let x and z be the usual quantities obtained from Table K–W, knowing q and let x' and z' be the values obtained knowing q'. Then, letting r be an abridged notation for r_t; the tetrachoric coefficient of correlation, or the correlation as found from a four-fold table assuming a normal correlation surface, is given by

$$\begin{aligned}\frac{\delta-qq'}{zz'}=r&+xx'\frac{r^2}{2!}+(x^2-1)(x'^2-1)\frac{r^3}{3!}+(x^3-3x)(x'^3-3x')\frac{r^4}{4!}\\&+(x^4-6x^2+3)(x'^4-6x'^2+3)\frac{r^5}{5!}\\&+(x^5-10x^3+15x)(x'^5-10x'^3+15x')\frac{r^6}{6!}\\&+(x^6-15x^4+45x^2-15)(x'^6-15x'^4+45x'^2-15)\frac{r^7}{7!}+\cdots\end{aligned}$$

(Equation giving r_t, the tetrachoric coefficient of correlation)..[207]

To express the law governing successive coefficients of powers of r let $v_n w_n/n$ be the coefficient of r^n, v_n be a function of x, and

w_n a function of x'; then v_n may be expressed in terms of v's of a lower order:

$$v_n = xv_{n-1} - (n-1)\,v_{n-2} \quad \text{and similarly} \quad w_n = x'w_{n-1} - (n-1)\,w_{n-2}$$
$$v_0 = 1,\ v_1 = x \quad \text{and similarly} \quad w_0 = 1,\ w_1 = x' \qquad [208]$$

Thus the equation as written to the r^7 term may be continued to any number of additional terms desired should it not converge rapidly enough to make terms above the r^7th negligible. For small values of r some slight simplification of the work will result from using Everitt's tables (1910). For values of r equal to or greater in absolute value than .80, tables (Everitt, 1912 and Lee, 1917) giving the δ for certain assigned r's and for various dichotomies are of great assistance, as they enable a determination of r by interpolation without the extensive labor involved in formula [207], or in Everitt's form of the same formula which utilizes his tables. The solution of equation [207] for r may follow the usual methods employed in the solution of a parabolic equation of higher degree than the second, but the method pursued in the following example is more expeditious for usual values of r. The data are extracted from the findings of the Division of Psychology of the Surgeon General's Office (Yerkes 1021, page 507).

TABLE XLIX

		Score on Army Intelligence Alpha Test		
		A or *B*	Below *B*	
First Lieutenants	Departments other than Medical	2940	431	3371
	Medical Department .	1799	590	2389
		4739	1021	5760

Same, expressed as proportions

$.5104 = \alpha$	$.0748 = \beta$	$.5852 = p$
$.3123 = \gamma$	$.1025 = \alpha$	$.4148 = q$
$.8227 = p'$	$.1773 = q'$	1.0000

Entering Table K–W with q we find

$$x = .215215$$
$$z = .389809$$

Entering with q' we find

$$x' = .925705$$
$$z' = .259914$$

Substituting these values in equation [207] we have

$$\frac{.1025 - .0735440}{.1013168} = r + .099613\, r^2 + .022741\, r^3 + .05255\, r^4 - .03195\, r^5 + .0288\, r^6 + \cdots$$

Solving the quadratic given by neglecting the last four terms, gives $r = .2781$. It is obvious by inspection of the signs of the terms neglected that this value is slightly too large. Let us therefore assume the value .2770, substitute it for r in the last five terms of the equation and solve for r to the first power for which we have not substituted. Doing so gives $r = .2773998$. The assumed value for r was too small. Let us therefore repeat the process, assuming $r = .2774$. This gives $r = .2773741$. We thus have the following table:

TABLE L

Assuming for Terms Involving Powers Other than the First that $r =$	Leads to $r =$
.2770	.2773998
.2774	.2773741

Interpolating between these two pairs of values so as to find that value starting with which leads to itself as result, we find $r = .2773757$. Expressed as an equation this value of r is given by

$$\frac{r - .2770}{.2774 - .2770} = \frac{.2773998 - r}{.2773998 - .2773741}.$$

The work has been carried to seven figures merely to show the method, not because such refinement in calculation is necessary in order to obtain a three or four figure result.

It will be noted that for this low correlation an excellent approximation, $r = .2781$ to the final answer, is obtained by keeping the first and second power terms only. We thus find the

correlation between being a lieutenant in the medical corps, as opposed to being one in some other corps, and low intelligence test standing to be .2774. We desire to know the probable error of this result. The full formula (Pearson 1900 cor.), is laborious to use and Pearson (1913 coef.) has given an equation which constitutes a close approximation to the full formula. We may give certain preliminary formulas. The first is Sheppard's:

$$r = \cos(2\pi\beta)$$ (Tetrachoric correlation in case both dichotomic lines are the medians). [209]

If the categories $(a + b)$ and $(a + c)$ correspond to positive deviations in the traits, then the measures represented by the a cell are $(+ +)$ measures, those by d $(- -)$, those by b $(+ -)$, and those by c $(- +)$ measures. Furthermore b must equal c so that $2\beta = \beta + \gamma$, – the proportion of unlike sign pairs. We may call this proportion u and write the preceding formula.

$$r = \cos(\pi u) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots [209\ a]$$

This very simple formula will give good results if the dichotomies differ slightly from the medians, but it should hardly be used if both p and p' are greater than .55, or if one is equal to .5 and the other greater than .6. The standard error of tetrachoric r when the dichotomies are at the medians is

$$\sigma_r = \frac{\sqrt{1 - r^2}}{\sqrt{N}} 2\pi\sqrt{\alpha\beta}$$ (The standard error of tetrachoric r when dichotomic lines are at the medians) . . . [210]

In case the true correlation is zero then no matter what the position of the dichotomic lines

$$\sigma_r = \frac{\sqrt{pqp'q'}}{zz'\sqrt{N}}$$ (The standard error of tetrachoric r when the real value of r = .00) . . . . . [211]

Finally when the true value of r is not zero, and when dichotomic lines are not at the medians, we have as a close approximation

$$\sigma_r = \frac{\sqrt{p\,p'qq'}}{\sqrt{N'zz'}} \sqrt{\left[1 - \left(\frac{\sin^{-1} r}{90^\circ}\right)^2\right](1 - r^2)}$$

(The general formula for the standard error of tetrachoric r) . . [212]

In tne reference cited are to be found tables of $\sqrt{pq}/z$ and of the radical function of r, which will expedite the calculation of the standard error. For the probable error of r we have

$$\text{P. E.}_r = \left\{.6457\sqrt{(1-r^2)\left[1-\left(\frac{\sin^{-1}r}{90^\circ}\right)^2\right]}\right\}\frac{\sqrt{pqp'q'}}{zz'\sqrt{N}}$$

(General formula for the probable error of tetrachoric r) . . [213]

The term in braces is tabled herewith.

TABLE LI

Functions Involved in Calculating the Probable Error of Tetrachoric r

r	Function of r	r	Function of r	r	Function of r
.00	.674	.60	.492	.80	.327
.10	.670	.61	.486	.81	.316
.20	.655	.62	.479	.82	.305
.25	.645	.63	.472	.83	.294
.30	.631	.64	.465	.84	.283
.35	.615	.65	.458	.85	.271
.40	.597	.66	.450	.86	.259
.42	.588	.67	.443	.87	.246
.44	.580	.68	.435	.88	.233
.46	.570	.69	.427	.89	.220
.48	.561	.70	.419	.90	.206
.50	.551	.71	.411	.91	.192
.51	.545	.72	.402	.92	.177
.52	.540	.73	.393	.93	.161
.53	.535	.74	.385	.94	.144
.54	.529	.75	.376	.95	.127
.55	.523	.76	.366	.96	.108
.56	.517	.77	.357	.97	.088
.57	.511	.78	.347	.98	.066
.58	.505	.79	.337	.99	.039
.59	.499			1.00	.000

We may use the preceding formula to calculate the probable error of the correlation between being a first lieutenant in the medical corps and low Army Alpha standing.

$$\text{P. E.}_r = \frac{.6378}{\sqrt{5760 \times .9397 \times .6661 \times 1.4656 \times .3158}} = .0156.$$

The item .6378 comes from Table LI; 5760 is the population; and the other items come from z/p and z/q columns of Table K–W.

Section 73. Correlation in a Four-fold Point Surface

In case the categories in a 2×2 fold table cannot reasonably be thought of as indicating different quantitative values of the variate, but of necessity as being indicative of qualitative differences, we may consider the distribution to be a point distribution, i.e., that the p frequencies are all concentrated at a single point and not spread over an interval, and similarly for q, p' and q'. It will make no difference what the numerical value of the difference between the two points of the distribution is, or in fact whether the value is, in the mathematical sense, real or imaginary. So we will call the distance between the p and q points j, and that between p' and q' points k, and calculate a regular product-moment coefficient of correlation using formula [93] and taking moments around the intersection of the p and p' category point values.

$$r = \frac{\delta jk - (qj)\,(q'k)}{\sqrt{qj^2 - (qj)^2}\sqrt{q'k^2 - (q'k)^2}} = \frac{\delta - qq'}{\sqrt{pq}\sqrt{p'q'}}$$

Algebraic transformation enables the writing of this formula in the form

$$\phi = r_{hk} = \frac{\alpha\delta - \beta\gamma}{\sqrt{pq\,p'q'}}$$

(Product-moment correlation between two point distributions. Pearson's r_{hk}; or ϕ, Yule's theoretical value of r) .....[214]

Pearson and Heron have called this coefficient the Boas-Yulean ϕ. For a discussion of it see Boas (Science, May 1, 1909, page 824), Yule (1912 meth.), and Pearson and Heron (1913). This formula may safely be used if the point nature of the distribution can be established. It would seem to be the appropriate formula in calculating the correlation between unit traits; possibly that, for example, between sex and albinism. The statistical criteria establishing the point nature of the value of a variate are still to be devised. They would constitute an important supplement to experimental and biological work. Pearson has shown (1900, con.) that r_{hk} (in the notation of this chapter and of Table K–W this is $r_{xx'}$) is the correlation between the means, if measured in terms of the standard deviations of their distributions, of two variates of a

2×2 fold normal correlation surface and that it is also (1904 theory), ϕ, the square root of the mean square contingency of a 2×2 fold table without any assumption of normality.

It is necessary to distinguish between r_{hk} and $r_{M_1M_2}$ of Section 49. This latter was found to equal r_{12} [formula 118]. But since

$$h = \frac{M_1}{\sigma_1} \text{ and } k = \frac{M_2}{\sigma_2}$$

it will be seen that only when division of the means by the standard deviations has no effect upon the correlation, would $r_{hk} = r_{12}$. This is not the case for continuous variates, so that ϕ or r_{hk} should not be taken as the correlation between continuous variates even if they are recorded in a two-category manner. The coefficient ϕ is a product-moment coefficient as concerns h and k or discrete variables, but with reference to continuous variables it belongs to group (2) which we will now consider.

Section 74. Measures of Correlation not Equivalent to the Product-Moment Coefficient; Yule's Coefficients of Association and of Colligation

Two coefficients developed by Yule may be considered in connection with ϕ. Using the same notation they are

$$Q = \frac{ad - bc}{ad + bc} \quad \text{(Yule's coefficient of association)} \ldots [215]$$

$$\omega = \frac{\sqrt{ad} - \sqrt{bc}}{\sqrt{ad} + \sqrt{bc}} \quad \text{(Yule's coefficient of colligation)} \ldots [216]$$

Yule (1912) points out that Q is not changed by multiplying the frequencies in the various categories. Thus the Q's for the two following tables, the second of which has been obtained from the first by multiplying the frequencies in the $(a + b)$ category by ten and those in the $(b + d)$ category by five, are identical.

a	b
c	d

$10\,a$	$50\,b$
c	$5\,d$

Yule claims this as a peculiar advantage of the coefficient, but for a coefficient to be stable under such violent treatment may

be looked upon as a detriment, as Pearson and Heron (1913) have shown. The coefficient of colligation has the value that ϕ takes when the 4-fold table is "equalized" and when the classes are given equal or their "natural" percentages to employ the term used by Yule. Thus given the 4-fold

a	b
c	d

let us multiply the first row, second row, first column and second column respectively by the fourth roots of the quantities

$$\frac{cd}{ab}, \quad \frac{ab}{cd}, \quad \frac{bd}{ac}, \quad \frac{ac}{bd}.$$

This gives the "equalized" 4-fold

$\sqrt{ad}$	$\sqrt{bc}$
$\sqrt{bc}$	$\sqrt{ad}$

in which plainly $p = q = p' = q' = .5$. The correlation ϕ may be calculated from this, noting that

$$\alpha = \delta = \frac{\sqrt{ad}}{N}, \ \beta = \gamma = \frac{\sqrt{bc}}{N}$$

so that

$$\phi = \frac{ad - bc}{\sqrt{(\sqrt{ad} + \sqrt{bc})^4}} = \frac{\sqrt{ad} - \sqrt{bc}}{\sqrt{ad} + \sqrt{bc}} = \omega.$$

Thus Yule's coefficient of colligation constitutes a ϕ calculated from the equalized table. Conditions which would warrant its use as a measure equivalent to a product-moment coefficient of correlation are seldom present. They are (*a*) point distribution in the traits and (*b*) warrant for equalization of the table. Warrant for equalizing may occasionally be present; as for example, if ten men and 100 women are measured and it is desired to find the correlation when the population of men and women are equal, but it is difficult to think of a reasonable problem in which there would be warrant for equalizing in the case of both traits. If ω has peculiar value, not as a product-moment coefficient, but as some other kind of a cor-

relation coefficient, its physical meaning is still to be demonstrated and meanwhile it would seem the part of wisdom to limit its use to the narrow field in which conditions (*a*) and (*b*) are met. A still narrower range of utility for the association coefficient Q seems indicated. The great ease with which Q and ω can be calculated, as compared with r_t and C, the contingency coefficient, will tempt one to use them for situations for which they are not applicable. Yule has derived the standard error of ϕ (see Pearson and Heron, 1913). It is

$$\sigma_\phi = \frac{1}{\sqrt{N}}\left\{1 - \phi^2 + \left(\phi + \frac{\phi^3}{2}\right)\left[\sqrt{\frac{q}{p}} - \sqrt{\frac{p}{q}}\right]\left[\sqrt{\frac{q'}{p'}} - \sqrt{\frac{p'}{q'}}\right] - \frac{3\phi^2}{4}\left(\frac{q}{p} + \frac{p}{q} - 2\right)\left(\frac{q'}{p'} + \frac{p'}{q'} - 2\right)\right\}^{\frac{1}{2}}$$

(Standard error of ϕ from a 4-fold table).....[217]

Although ω is a special case of ϕ, the multiplication of the frequencies to obtain the equalized 4-fold table introduces another factor so that we cannot in general take σ_ω as being equal to σ_ϕ.

The contingency method developed by Pearson leads to two constants. One is P, the probability of a situation as extreme, as that found, arising as a matter of chance if the two variables are in truth uncorrelated; hence if P is small it argues for a correlation. The second is C, the coefficient of contingency, which under certain conditions is equal to the coefficient of correlation which would be obtained from the same data. The coefficient of contingency belongs to the first group of measures of relationship, but as it is derived in connection with P we will consider it here.

Section 75. Measures of Relationship Interpreted in Terms of Probability

The product theorem in probabilities is that if p is the probability of occurrence of a certain event and p' of a second unrelated event, then pp' is the probability of the joint occurrence of the two events. Thus if 30 per cent of a given population have blue eyes and if 50 per cent are males and if eye color and sex are uncorrelated, then the likelihood, in making a random selection of obtaining a blue-eyed male, is .15 (= .30 × .50) or, in the long run, 15 per cent of the random selections

will be blue-eyed males. If, then, a large drawing results in a proportion sufficiently different from .15 to preclude the possibility of chance, the existence of correlation between eye color and sex is established. We need to know P, as defined in the last paragraph of Section 74, and we desire a measure based upon P which is comparable in its general meaning to a product-moment coefficient of correlation. Let us be given the manifold table.

TABLE LII

n_{1a}	n_{1b}	n_{1c}	n_1
n_{2a}	n_{2b}	n_{2c}	n_2
n_{3a}	n_{3b}	n_{3c}	n_3
n_{4a}	n_{4b}	n_{4c}	n_4
n_a	n_b	n_c	N

designated n_s

designated $n_{s'}$

in which n is the number of cases in a category of the first variable, $n_{s'}$ in a category of the second, and $n_{ss'}$ the number in the cell given by the intersection of the n_s and $n_{s'}$ categories. There are as many n_s frequencies as there are categories in the first variable, as many $n_{s'}$ frequencies as categories in the second variable and as many $n_{ss'}$ frequencies as the product of the number of categories in the first variable times the number in the second. If a chance situation maintains, the proportion of the whole found in a cell will, by the product theorem, be given by

$$\frac{n_{ss'}}{N} = \frac{n_s}{N} \cdot \frac{n_{s'}}{N} \text{ or } n_{ss'} = \frac{n_s n_{s'}}{N} \quad \ldots\ldots\ldots\ldots\ldots\ldots [218]$$

In general this situation will not maintain, so that the actual number in a compartment minus the chance or theoretical number, measures the divergence of the situation from chance. This magnitude will be designated by $d_{ss'}$ and will be called the cell divergence

$$d_{ss'} = n_{ss'} - \frac{n_s n_{s'}}{N} \quad \text{(Cell divergence from chance situation)} \ldots\ldots\ldots\ldots\ldots [219]$$

The cell divergence is the divergence of a cell frequency from a chance frequency when it is desired to compare the obtained

situation with the uncorrelated or chance situation; but if it is desired to test out some theoretical cell frequencies ($m_{ss'}$), then the cell divergence becomes the divergence of an actual cell frequency from the theoretical frequency, or ($n_{ss'} - m_{ss'}$). Therefore we have for the general case

$$d_{ss'} = n_{ss'} - m_{ss'} \quad \text{(The cell divergence)} \ldots\ldots\ldots [220]$$

The square of the cell divergence divided by the theoretical frequency (which is usually the chance frequency) will be called the cell square contingency, while the sum of all such cell square contingencies has been termed by Pearson (1900, crit.) the square contingency, and given the symbol χ^2. Thus

$$\chi^2 = S\left(\frac{d^2_{ss'}}{m_{ss'}}\right) = S\left(\frac{(n_{ss'} - m_{ss'})^2}{m_{ss'}}\right) \quad \text{(The square contingency)} \ldots\ldots [221]$$

Obviously

$$S\, d_{ss'} = 0.$$

A measure of total contingency can be built upon the absolute values of the cell divergencies, $| d_{ss'} |$ (Pearson, 1904), but the measure of square contingency has superior advantages.

The square contingency cannot be directly interpreted because two factors are involved in it, the number of cells and the strength of the contingency. To eliminate the number of cells from consideration, Pearson has given the two equations

$$P = \sqrt{\frac{2}{\pi}}\int_{\chi}^{\infty} e^{-\frac{1}{2}\chi^2} d\chi + \sqrt{\frac{2}{\pi}} e^{-\frac{1}{2}\chi^2}\left(\frac{\chi}{1} + \frac{\chi^3}{1\cdot 3} + \frac{\chi^5}{1\cdot 3\cdot 5} + \cdots + \frac{\chi^{n'-3}}{1\cdot 3\cdot 5\cdots(n'-3)}\right) \quad \text{if } n' \text{ be even} \ldots [222]$$

$$P = e^{-\frac{1}{2}\chi^2}\left(1 + \frac{\chi^2}{2} + \frac{\chi^4}{2\cdot 4} + \frac{\chi^6}{2\cdot 4\cdot 6} + \cdots + \frac{\chi^{n'-3}}{2\cdot 4\cdot 6\cdots(n'-3)}\right) \quad \text{if } n' \text{ be odd} \ldots\ldots [222\,a]$$

in which n' is the number of cells and P the probability that random sampling would lead to as large or larger divergence between theory and observation. Elderton (1902 tables, and also, Pearson's Tables) has tabled P for various numbers of cells and values of χ^2. It is thus a simple matter to determine the probability of a situation as extreme as the one observed (note that this is not equivalent to saying "the probability of the observed situation") arising as a matter of chance. There is no assumption of normality in the determination of χ^2, but

in deriving the equation giving P from χ^2 it is assumed that the frequencies in a cell resulting from successive samplings form a normal system of variates. This is entirely different from the assumption that the categories are classes in reality constituting a normal distribution. It is because of avoiding any such assumption that the contingency method has its chief value. The assumption that, within a single cell, the results of successive samplings will constitute a normal distribution of frequencies, may regularly be expected to hold, provided p, the probability of a measure being in a cell, is not so small but that $(p + q)^n$ can be approximately represented by a normal distribution. As a practical matter $(p\,N)$ the theoretical number of cases in the cell should not be less than 1.00. If the categories are such that the theoretical frequency in any cell is less than 1.00, two or more categories should be combined so as to give cells with theoretical frequencies greater than 1.00. As a very minimum, not to be approached if avoidable, the smallest theoretical frequencies should not be less than .7.

Section 76. Equi-probable r

In case p is very small, its meaning is difficult to interpret. Pearson (1912 novel) has pointed out that the improbability of the obtained 4-fold arising as a matter of chance is equal to the improbability of a tetrachoric coefficient of correlation of a certain magnitude based upon the same number of cases, and Pearson and Bell have provided tables (see Pearson's tables) whereby a P calculated from a 4-fold table may be used to determine an equally improbable tetrachoric coefficient of correlation. Pearson does not recommend this method of interpreting P in case of extreme dichotomies, or in any case as being preferable to tetrachoric r.

Section 77. Mean Square Contingency and Coefficient of Contingency

We have obtained a measure of probability, P, from the square contingency χ^2. We may also interpret the results by means of a coefficient of contingency. The most valuable form is that derived by Pearson which he has called C_2 and

which we will here call C. We will first need the mean square contingency. Designating it by ϕ^2 we have

$$\phi^2 = \frac{\chi^2}{N} \quad \text{(Mean square contingency)} . [223]$$

The magnitude ϕ as thus defined is identical in the case of a 4-fold table with ϕ of formula [214]. As here defined it is obtained from a manifold of any number of cells. As has been pointed out in the case of a 4-fold table, ϕ is not a coefficient of correlation of a graduated or continuous variate, nor is the function

$$C = \sqrt{\frac{\phi^2}{1 + \phi^2}} = \sqrt{\frac{\chi^2}{N + \chi^2}} \quad \text{(Coefficient of contingency)} . [224]$$

but the latter is comparable with it. In fact, if for each variable the categories are successive values of a graduated variate, and if the population is large and the number of categories great so that there is not a grouping error, and if the correlation surface is normal, then C is identical with the product-moment coefficient of correlation.

As a measure of relationship between continuous variates there are two corrections which should be applied to C, one due to number of cells and the other a correction for class index. (Pearson and Heron 1913, page 217.)

If κ = number of rows and λ = the number of columns and if the frequencies in the categories do not differ one from another in an extreme manner, the corrected mean square contingency, ${}_c\phi^2$, is given by

$${}_c\phi^2 = \frac{\chi^2 - (\kappa - 1)(\lambda - 1)}{N} \quad \text{(Value of } \phi^2 \text{ corrected for number of cells)} . [225]$$

In case broad categories are used there are wide differences in the measures within a category and these may be differently grouped for the successive cells of a single category, so that there is a correction for class index needed (Pearson 1913, meas., page 130). This correction does not apply to ϕ which is, in the case of a 4-fold, the correlation between points and may be thought of as a similar sort of a function in the case of a manifold of a greater number of cells; but it does apply to C, the coefficient of contingency, which aims to measure the relationship between continuous or graduated variates. Thus

we will consider the uncorrected C as the correlation between class means and correct by formula [103] where $r_{\chi x}$ and $_{\gamma y}$ have the meanings defined by formula [102]. The student must not confuse χ of formulas [192], [103], [226] and [226 *a*] with the mean square contingency, χ^2 of formula [224]. They are entirely unrelated. Applying the correction, we have

$$mC = \frac{C}{r_{\chi x} r_{\gamma y}} \quad \text{(Coefficient of contingency corrected for class means)} \ldots\ldots [226]$$

We must now obtain values for the correlation between the variates and the class means, $r_{\chi x}$ and $r_{\gamma y}$. The preceding formula may be written in a form similar to formula [103].

$$mC = \frac{C}{\sigma_x \sigma_y} \quad \ldots\ldots [226\ b]$$

Note that the assumption of normality implies that the corrective factor $1/(\sigma_x \sigma_y)$ is as great for the problem in hand as it would be were the distribution of the two traits normal. In other words we assume normality only in the problem of determining the corrective factor and not in the determining of C. Wide divergencies from normality would probably amount to very little so far as the corrective factor is concerned, and as it is necessary to make some assumption in order to determine this factor we can do no better than assume normality of distribution. Doing this we find σ_x and σ_y as was done in Section 47. Should we not wish to make the assumption of normality we may assume a rectangular distribution and find the correlation between class means and variates. A rectangular distribution of κ units length has a standard deviation of $\sqrt{\kappa^2/12}$ and the standard deviation of the means of a rectangular distribution κ units in length divided into κ equal intervals is $\sqrt{\kappa(\kappa - 1)/12}$. Thus the corrective factor is determined from

$$r_{x\chi} = \frac{\sigma_\chi}{\sigma_x} = \sqrt{\frac{\kappa}{\kappa - 1}}$$

(Correlation between class means and variates assuming a rectangular distribution of κ equal sub-ranges) $\ldots\ldots [227]$

If κ is the number of categories in the first variable and λ the number in the second, the total corrective factor is

$$\sqrt{\frac{\kappa\lambda}{(\kappa - 1)(\lambda - 1)}} \quad \ldots\ldots [228]$$

This correction is larger than the one based upon the assumption of normality and probably is in general less sound. The following table is given to show the magnitude of the corrective factors upon various assumptions and to provide $r_{\chi x}$ when certain assumptions are reasonable without entailing the detailed calculation.

TABLE LIII

Value of $r_{\chi x}$, the Correlation between the Class Means and Variates for Different Groupings

Number of Classes	Equal Ranges. Normal Distribution	Equal Sub-Frequencies. Normal Distribution	Equal Ranges. Rectangular Distribution	Equal Ranges. Any Distribution
2	.798	.798	.707	.589
3	.872	.891	.816	.842
4	.923	.928	.866	.915
5	.949	.947	.894	.946
6	.964	.959	.913	.963
8	.979	.972	.935	.979
10	.986	.979	.949	.987
15	.993	.988	.966	.994
20	.996	.992	.975	.997

The values in the last column have been derived upon the assumption that a parabola would well represent the frequency surface of any three neighboring classes. In the calculation of the first and last columns of this table it has been assumed that the total range was equal to 5.6 standard deviations which would approximately be the case in a normal distribution if the total population is 100 (see prob. 1, Chapter V). Pearson (1913 inf.) gives a table containing in part similar information upon the assumption that the total range equals 6.0 standard deviations which is approximately the case if the total population is 185. The corrective factors given in the 1st, 2d, and 4th columns are nearly equal to each other if the number of classes is greater than three, so it makes little difference which of these three hypotheses is assumed in determining this corrective factor. The assumption of a rectangular distribution leads to quite different results throughout the entire length of the table.

We have considered two corrections, one the correction of ϕ for number of cells and the second a correction of C for use of class means instead of variates in the classes. One further

important item is the probable error of the contingency coefficient. Much study of this point has been made (Blakeman and Pearson 1906 and Pearson 1915, prob.) and certain of the methods obtained are involved. The method here given, derived by Pearson (1915 prob.), is fairly simple, involving the calculation of but a single additional constant ψ^3. Let the cell ψ^3 function be defined by the equation

$$\text{Cell } \psi^3 \text{ function} = \frac{(d_{ss'})^3}{(m_{ss'})^2} \quad \ldots\ldots\ldots\ldots\ldots\ldots [229]$$

and let ψ^3 be the sum of such functions for the entire table, divided by the population, thus

$$\psi^3 = \frac{1}{N} S\left(\frac{(d_{ss'})^3}{(m_{ss'})^2}\right) \quad (\psi^3 \text{ function required in finding the probable error of } \phi \text{ and of } C) \ldots [230]$$

Having ϕ^2 and ψ^3 we may obtain the standard error of ϕ from the formula

$$\sigma_\phi = \frac{1}{\sqrt{N}}\left(\frac{\psi^3}{\phi^2} + 1 - \phi^2\right)^{\frac{1}{2}} \quad \text{(Standard error of } \phi) \ldots\ldots [231]$$

Further, having C we obtain

$$\sigma_C = \frac{1}{\sqrt{N}}\left(\frac{\frac{\psi^3}{\phi^2} + 1 - \phi^2}{(1+\phi^2)^3}\right)^{\frac{1}{2}} \quad \text{(Standard error of the coefficient of mean square contingency)} \ldots\ldots\ldots [232]$$

We may illustrate the calculation of ϕ, C, σ_c and the corrections to ϕ and C by the following data taken from the army psychological findings (Yerkes, 1921, page 825).

TABLE LIV

	Baker	Band Musician	Barber	Bookkeeper	Butcher	
Tested by Army Alpha	294.	289.	275.	450.	370.	1678
	323.5	262.9	321.8	390.9	378.9	
	− 29.5	26.1	− 46.8	59.1	− 8.9	
	2.690	2.591	6.821	8.935	.209	
	− .245	.257	− .992	1.351	− .005	
Tested by Army Beta	85.	19.	102.	8.	74.	288
	55.5	45.1	55.2	67.1	65.1	
	29.5	− 26.1	46.8	− 59.1	8.9	
	15.680	15.104	39.678	52.054	1.217	
	8.334	− 8.741	33.640	− 45.848	.167	
	379	308	377	458	444	1966

$\chi^2 = 144.979$ $\quad N\psi^3 = -12.082$

$\phi^2 = .07374$ $\quad \psi^3 = -.006145$

$C = .2621$ $\quad \sigma_c = .01861$

P. E.$_c$ = .0126

The 5 entries in each cell are, in order, as follows:

$n_{ss'}$	The frequency found in the cell
$m_{ss'}$	The theoretical cell frequency
$d_{ss'}$	The cell divergence
$\frac{(d_{ss'})^2}{m_{ss'}}$	The cell square contingency
$\frac{(d_{ss'})^2}{m_{ss'}} \times \frac{d_{ss'}}{m_{ss'}}$	The cell ψ^3 function

It should be noted that the ϕ^2 used in the calculation of σ_c is not the corrected value. We may, however, with insignificant error consider σ_c to be either the standard error of the raw or the corrected coefficient of contingency. The mean square contingency corrected for too fine grouping, ${}_c\phi^2$, is, by formula [225],

$$ {}_c\phi^2 = \phi^2 - \frac{(\kappa - 1)(\lambda - 1)}{N} $$

$$ = .07374 - \frac{4 \times 1}{1966} = .07171 $$

The corrected coefficient of contingency depends upon the correlation between class means and variates. Let $r_{\chi x}$ stand for this correlation in the case of the test series and let $r_{\gamma y}$ be this correlation for the vocation series. It is very difficult to make an assumption as to the distribution of the variates within the vocational categories. However, assuming "equal ranges any type of frequency" we find from Table LIII that $r_{\gamma y} = .946$, for a five category series. The assumption of a normal distribution for the other variable is reasonable though we cannot expect the most reasonable of assumptions to give a very reliable corrective factor from a two category distribution. We have

TABLE LV

	Number	Per Cent	z	$\bar{x}$ Mean of Class
Test by Army Alpha	1678	85.36	.2294	.257
Test by Army Beta	288	14.64		− 1.567

$$ \sigma_x = \sqrt{8536\ (.257)^2 + .1464\ (-\ 1.567)^2} = .6449 - $$

and since $\sigma_{\chi'} = 1.00$ we have

$$r_{x\chi} = \frac{.645}{1.000} = .645$$

Thus finally

$${}_mC = \frac{\sqrt{\frac{{}_c\phi^2}{1 + {}_c\phi^2}}}{r_{x\chi} r_{y\gamma}} = \frac{.2587}{.6449 \times .946} = .4240$$

P. E. of ${}_mC$ = approximately .0204 as determined from the proportion

$$\frac{.2621}{.0126} = \frac{.4240}{\text{P. E. of } {}_mC}$$

This completes the solution, and for the problem in hand we may conclude that there is a small correlation of .424 between trades considered and literacy and that this is established with a very satisfactory degree of certainty.

The reader should note that the corrected value of c differs materially from the raw value.

Section 78. Variate Difference Method

The variate difference method was first used by Miss F. E. Cave, in 1904, in a study of the correlation of barometric heights, published in the proceedings of the Royal Society of London, v. 74, pp. 407. The object of this study was to get rid of seasonal change by correlating first differences of readings as obtained at two stations. Later, Hooker (1905, Jour. of the Roy. Soc., v. 68), Student (1914), Anderson (1914), Beatrice M. Cave and Pearson (1914) and Ritchie-Scott (1915) have further developed the theory and illustrated its use, and Persons (1916), (1917) has noted certain of its shortcomings. There is still much to be done in establishing its degree of applicability to short series such as are usually available in material influenced by spurious time and space factors.

If barometric heights constitute the data and a large number of measures are available, there is little doubt but that the method will give the correlation between the readings at two stations independent of spurious space or time factors; but if two series of yearly price indexes, extending over n years,

where n is small, are correlated by the variate difference method; (*a*) the probable error of the correlation obtained is not definitely known, (*b*) the number of differences which it is desirable to use is uncertain, and (*c*) the relation between the applicability of the method and the size of n is not established. Cave and Pearson (1914) consider good results to be obtained by going to fifth or sixth order differences when dealing with eleven commercial indexes, each extending over 28 years, but this point is not indubitably established. The problem shortly to be presented to illustrate the method is equally extensive in time, but the real relationship between the variables, independent of time, can hardly be said to be apparent. The treatment of the following sections will be in the order, (*a*) notation, and tests of applicability, [1] by comparison of standard deviations of successive difference series and [2] by the stability of the successively obtained correlations; and (*b*) illustration by a problem.

(*a*) Given two series, $x_1, x_2, \cdots x_n$ and $y_1, y_2, \cdots y_n$ between which there is an organic correlation, R, and a spurious correlation due to a time or location factor such that the two phenomena together result in an apparent, i.e., an obtained correlation, of r. The problem is to determine R. Student (1914) has shown that if

$$\left.\begin{aligned} x_1 &= X_1 + bt_1 + ct^2_1 + dt^3_1 + \text{etc.} \\ x_2 &= X_2 + bt_2 + ct^2_2 + dt^3_2 + \text{etc.} \\ &\text{etc.} \end{aligned}\right\} \quad \ldots\ldots\ldots\ldots [233]$$

and if

$$\left.\begin{aligned} y_1 &= Y_1 + Bt_1 + Ct^2_1 + Dt^3_1 + \text{etc.} \\ y_2 &= Y_2 + Bt_2 + Ct^2_2 + Dt^3_2 + \text{etc.} \\ &\text{etc.} \end{aligned}\right\} \quad \ldots\ldots\ldots [234]$$

in which X_1, X_2, etc., Y_1, Y_2, etc., are independent of time or location, then, if the parabolic equations in t terminate with some power t^n, the correlation r_{XY} is given by the correlation between Δ_n and δ_n, the two series of n-th order differences, Δ_1 standing for the measures $(x_1 - x_2)$, $(x_2 - x_3) \ldots (x_{n-1} - x_n)$; Δ_2 for the measures $[(x_1 - x_2) - (x_2 - x_3)]$, $[(x_2 - x_3) - (x_3 - x_4)], \ldots [(x_{n-2} - x_{n-1}) - (x_{n-1} - x_n)]$; and similarly Δ_3 for third order differences; Δ_4 for fourth order

differences; etc.; the δ's having comparable meanings in the case of the y-series. Cave and Pearson have noted that in this equation the ratio

$$\frac{\sigma_{\Delta_m}}{\sigma_{\Delta_{m+1}}} = 4 - \frac{2}{m+1} \quad \text{.....[235]}$$

and that, therefore, starting with a series in which measures are not independent but influenced by a time factor which can be expressed, as suggested, by a terminating parabolic series, taking successive differences and calculating the standard deviations of the difference series, one should obtain, as soon as sufficient differences have been taken to eliminate the spurious time factor, standard deviations bearing the ratio indicated. This accordingly constitutes a test in a single series of the number of differences which are required to eliminate a time or space factor. Cave and Pearson applied this test to the eleven series with which they worked, but did not succeed in establishing the number of differences necessary to eliminate the time factor. They attribute their failure to the small period studied. However, 28 years is, as economic data run, a fairly long period. Some method, — partial correlation, variate difference, or what not, — to eliminate an annoying time factor, for data covering such or a shorter period, is greatly needed.

The approach of the ratio of successive standard deviations of the difference series of the single variable to $4 - 2/(m+1)$ is the first test of the possibility of eliminating a time or space factor by dealing with differences.

The second test lies in the stability of successive correlations between differences, of equal order, of the two series. Thus, if $r_{xy} \neq r_{\Delta_1\delta_1} \neq r_{\Delta_2\delta_2}$, but, very approximately, $r_{\Delta_2\delta_2} = r_{\Delta_3\delta_3} = r_{\Delta_4\delta_4}$, it would be concluded that the time or space factor had been eliminated by the resort to second differences and that the correlation then found, $r_{\Delta_2\delta_2}$ was in truth r_{XY}, the desired correlation between the two traits independent of the spurious element.

The data in Table LVI, p. 274, kindly supplied by Mr. Willis H. Rich, have all the characteristics expected in series to be treated by this method. That the conclusions will be found

to be somewhat doubtful points the weakness of the method in its present state of development.

TABLE LVI

Chinook Salmon — Columbia River

Date of Pack	Pack in 1000's of Cases	Hatchery Output in Millions of Fry	Fry Liberated in Spring of
1889	265		
1890	335		
1891	353		
1892	344		
1893	288	2.77	1890
1894	351	4.90	1891
1895	444	1.33	1892
1896	370	4.10	1893
1897	442	.21	1894
1898	346	.00	1895
1899	286	3.39	1896
1900	294	6.59	1897
1901	334 (1)	21.94	1898
1902	375	12.87	1899
1903	469	11.00	1900
1904	547	10.04	1901
1905	572	24.10	1902
1906	511	20.44	1903
1907	410	23.56	1904
1908	334	9.15	1905
1909	300	17.13	1906
1910	442	9.10	1907
1911	609	16.44	1908
1912	365	15.43	1909
1913	335	12.54	1910
1914	419	13.97 (2)	1911
1915	508	15.41	1912
1916	511	26.10	1913
1917	450	41.58	1914
1918	445	44.45	1915
1919	475	53.24	1916
1920	477	25.03	1917
		56.80	1918
		22.57	1919
		25.00 (3)	1920

(1) 334 is an estimate based upon the total pack for the year.
(2) 13.97 is an estimate based on the total hatchery output.
(3) 25.00 is a sheer estimate.

The problem is to ascertain if there is positive correlation between the number of fry liberated from the hatcheries and the run of salmon later, particularly three years later, when the fry are grown and return to spawn. It is known that the salmon returns to the same river in which liberated and that roughly 8 per cent (these 8 per cent are small fish and would be equivalent to some 5 per cent in weight of pack) return to spawn one year after liberation, 20 per cent (or 15 per cent of the pack) return in two years, 50 per cent (or 50 per cent of the pack) return in three years, 20 per cent (or 25 per cent of the pack) return in four years, and 2 per cent (or 5 per cent of the pack) return in five years. Accordingly if there is positive correlation between number of fry liberated and size of pack independent of time, it should be greatest when correlating size of pack with number of fry liberated three years earlier.

The means and squared standard deviations are given in the first two columns of accompanying Table LVII.

TABLE LVII

Means	Standard Deviations Squared	Ratios		Ratios of Ratios
		$\frac{\sigma^2_{m+1}}{\sigma^2_m}$	$4 - \frac{2}{m+1}$	
x 418.18	7,731.14			
		.971	2.000	.486
Δ_1 − 7.00	7,505.19			
		1.939	3.000	.646
Δ_2 − 2.35	14,552.54			
		2.660	3.333	.798
Δ_3 2.35	38,704.94			
		3.077	3.500	.879
Δ_4 5.58	119,100.91			
		3.267	3.600	.908
Δ_5 24.00	389,056.35			
		3.327	3.667	.907
Δ_6 56.18	1,294,512.40			

The last column of the table shows the approach of the ratios of the standard deviations squared to a random situation, i.e., a situation from which the time or space factor has been eliminated. There is seen to be some approach to the value $4 - 2/(m + 1)$, but the approach is not sufficiently close to say that this test supports the contention that a resort to fourth, fifth, or sixth differences frees the data of the spurious factor.

More promising results are obtained from the "hatchery

output" data. Keeping the data to the nearest .1 and shifting the decimal point one place to the right it yields

TABLE LVIII

	Standard Deviations Squared	Ratios		Ratio of Ratios
		$\frac{\sigma^2_{m+1}}{\sigma^2_m}$	$4 - \frac{2}{m+1}$	
y 159.50	17,131.96			
δ_1 — 8.22	8,231.07	.480	2.000	.240
δ_2 — 11.65	17,507.46	2.127	3.000	.709
δ_3 12.52	50,770.33	2.900	3.333	.870
δ_4 — 22.96	178,983.22	3.525	3.500	1.007
δ_5 15.87	674,219.57	3.767	3.600	1.046
δ_6 — 66.64	2,599,756.51	3.856	3.667	1.052

We may conclude that so far as this test permits us to form a judgment we will succeed in eliminating the spurious factor by resorting to fourth or higher differences.

Calculation of the product-moment coefficients of correlation between similar difference series gives the values recorded in the following table:

TABLE LIX

$$r_{xy} = .3802 \pm .1275$$
$$r_{\Delta_1\delta_1} = .0003 \pm .1580$$
$$r_{\Delta_2\delta_2} = .0145 \pm .1826$$
$$r_{\Delta_3\delta_3} = -.0258 \pm .2023$$
$$r_{\Delta_4\delta_4} = -.0247 \pm .2196$$
$$r_{\Delta_5\delta_5} = -.0005 \pm .2354$$
$$r_{\Delta_6\delta_6} = .0525 \pm .2503$$

The probable errors have been calculated by the following formulas, which are due to Anderson (1914): Let r_{XY} be, as before, the correlation between the two variables independent of the time or location factor; let σ_{00} be the standard error of r_{xy}; σ_{11} the standard error of $r_{\Delta_1\delta_1}$; σ_{22} the standard error of $r_{\Delta_2\delta_2}$, etc. Then,

(Standard errors of variate difference correlation coefficients) .....[236]

$$\sigma_{00} = \frac{1 - r^2_{XY}}{\sqrt{N}}$$

$$\sigma_{11} = \frac{1 - r^2_{XY}}{N - 1}\sqrt{\frac{3N - 4}{2}}$$

$$\sigma_{22} = \frac{1 - r^2_{XY}}{N - 2}\sqrt{\frac{35N - 88}{18}}$$

$$\sigma_{33} = \frac{1 - r^2_{XY}}{N - 3}\sqrt{\frac{231\,N - 843}{100}}$$

$$\sigma_{44} = \frac{1 - r^2_{XY}}{N - 4}\sqrt{\frac{1287\,N - 6128}{490}}$$

$$\sigma_{55} = \frac{1 - r^2_{XY}}{N - 5}\sqrt{\frac{46189\,N - 270635}{15876}}$$

$$\sigma_{66} = \frac{1 - r^2_{XY}}{N - 6}\sqrt{\frac{676039\,N - 4696566}{213444}}$$

. . . . . .

$$\sigma_{kk} = \frac{1 - r^2_{XY}}{N - k}\Bigg((N - k) + 2\,(N - k - 1)\left[\frac{k}{k + 1}\right]^2 + 2\,(N - k - 2)\left[\frac{k\,(k - 1)}{(k + 1)\,(k + 2)}\right]^2 + 2\,(N - k - 3)\left[\frac{k\,(k - 1)\,(k - 2)}{(k + 1)\,(k + 2)\,(k + 3)}\right]^2 + \cdots\Bigg)^{\frac{1}{2}}$$

The N throughout the formulas is the original population and not the reduced number of differences. The final correlation, r_{XY}, which maintains after elimination of the spurious factor, enters into all of these formulas. This correlation is of course not known, but if successive difference correlations remain approximately equal one may take this constant value as the value of r_{XY} and determine approximate probable errors. For the problem in hand we see that the first, second, third, fourth, fifth and sixth difference correlations are closely equal to zero. Accordingly, taking zero as the value of r_{XY} and using formula [236] we obtain the probable errors listed. Note that the standard error of r_{xy} is given as $(1 - r^2_{XY})/\sqrt{N}$ and not the usual value $(1 - r^2_{xy})/\sqrt{N}$. That is to say, r_{xy}, could it be assumed to be a measure of r_{XY}, has the standard error $(1 - r^2_{XY})/\sqrt{N}$, but as a measure not distinct from the space or time factor it has the usual standard error. In our present problem, since $r_{xy}/r_{\Delta_1\delta_1}$ does not approximately = 1.00 we should not assume it to be a measure of r_{XY}.

The conclusion which this treatment suggests is that there is no relation between planting of fry and run of salmon three years later, but this is in no sense established, due to the large probable errors. It is of course unfortunate that, with the very type of data for which large populations cannot be secured, the probable errors should be larger than for straight correla-

tions. This is a weakness of the method in the field for which it would otherwise be most serviceable.

It would be valuable to compare at length results obtained by the variate difference method with those from a partial correlation or partial correlation ratio method. The data in hand do not warrant too detailed an analysis, but it may be stated that, assuming either a rectilinear or a single flexion curvilinear regression line between time and each of the other two variables, the partial correlation between number of fry liberated and run three years later is positive and slightly greater than its probable error. Thus, for these data, the two methods do not point in the same direction.

Calculating variate difference correlation coefficients between number of fry liberated and run two, and again four, years later yield equally inconclusive results with those reported.

CHAPTER XI

MULTIPLE CORRELATION

Section 79. The Problem

The fundamental problem of multiple correlation is the estimation, with minimal error, of one variable knowing several others. Thus if X_0 is the dependent variable, or the one to be estimated, and X_1, $X_2 \cdots$, X_n the independent variables, and if $\overline{X}_0$ is the value of the dependent variable as estimated from the known X_1, $X_2, \cdots X_n$ variables, we may write

$$\overline{X}_0 = f(X_1, X_2, \cdots X_n)$$

and we will say that that function which makes

$$\frac{\Sigma (X_0 - \overline{X}_0)^2}{N} = \text{a minimum} \quad \ldots\ldots\ldots\ldots\ldots [237]$$

is the best function. Since $(X_0 - \overline{X}_0)$ is an error of estimate, this is identical with imposing the condition that the sum of the squares of the errors of estimate shall be a minimum. Just as we have found that there are many methods of measuring correlation, so there are many ways of measuring multiple correlation. The five following are important, but not inclusive of all possible methods.

(*a*) When $f(X_1 X_2 \cdots X_n)$ is a linear function of the variables we have the usual multiple correlation problem, and the method to be used is both the simplest and the most readily interpreted.

(*b*) When f is a known, but non-rectilinear function of the X's, appropriate transformations as suggested in Section 52 will ordinarily enable the treatment of this problem by methods applicable to (*a*).

The complete problem of simple or multiple correlation involves, as has been stated, (1) a measure of the strength of

relationship between a dependent variable and one or more independent variables, and also (2) an algebraic means of estimating the dependent variable knowing the independent variable, or variables. Whereas methods (*a*) and (*b*) preceding give solutions of both (1) and (2), methods (*c*), (*d*) and (*e*) following provide a solution of (1) only.

(*c*) A multiple and partial correlation ratio method enabling an estimation of the magnitudes of the multiple and partial correlations between graduated variables which are not related to each other by means of rectilinear regression lines. Also, a

(*d*) Multiple and partial contingency method accomplishing the same result as multiple and partial correlation ratios, and particularly applicable to data recorded in a categorical manner. This method also leads to interpretation in terms of probability.

(*e*) The variate difference correlation method. This method is of service when a time or space factor not showing rectilinear relation with the other two variables involved hides or clouds the partial relationships between the two variables. This method has been presented in the preceding section and is very different from (*a*), (*b*) and (*d*). The treatment of the next five sections is confined to method (*a*) and covers the 3 or 4 variable problem in Sections 80, 81, 82, the 4, 5, or 6-variable problem in Section 83, and the many variable problem in Section 84.

Section 80. Theoretical Treatment — 3 Variables

A simple three variable problem, so chosen that the interpretation is not complicated by unequal variabilities of the three series, will show the concrete and tangible significance of the partial and multiple correlation coefficients.

We shall use the following notation.

X = a gross score.

$x = X - M$ = a score as a deviation from the mean.

$\sigma = \sigma_x = \sigma_X =$ the standard deviation of either the x's or the X's

$$z = \frac{x}{\sigma} = \frac{X - M}{\sigma} = \text{a standard measure}$$

$$\sigma^2{}_z = \frac{\Sigma x^2}{N\sigma^2} = 1.0$$

Symbols with subscript zero as X_0, x_0, σ_0, z_0, designate the criterion or dependent variable. Symbols with subscript 1 designate the first independent variable, with subscript 2 the second independent variable. The following symbols with superior bars $\overline{X}_0$, $\bar{x}_0$, $\bar{z}_0$, designate gross criterion scores estimated from a knowledge of the independent variables, deviation scores estimated from such a knowledge, and standard scores estimated from such a knowledge, respectively. The statistical problem is to determine the two constants $\beta_{01\cdot 2}$ and $\beta_{02\cdot 1}$ (the significance of the subscripts is explained later) in the equation

$$\bar{z}_0 = \beta_{01\cdot 2} z_1 + \beta_{02\cdot 1} z_2 \quad \text{(Fundamental regression equation connecting standard measures — 3 variables)} \ldots\ldots\ldots [238]$$

so that the standard error of estimate $k_{0\cdot 12}$ is a minimum.

$$(z_0 - \bar{z}_0) \quad \text{(Error of estimate or residual of a standard criterion measure)} \ [239]$$

is the difference between the actual standard criterion score and the criterion standard score estimated from the independent variables. It is thus an error of estimate and the standard error of estimate is

$$k_{0\cdot 12} = \sqrt{\frac{\Sigma (z_0 - \bar{z}_0)^2}{N}} \quad \text{(Standard error of estimate of the standard criterion measures)} \ldots [240]$$

If z_1 and z_2 are worthless in shedding light upon the value of z_0 then $\beta_{01\cdot 2}$ and $\beta_{02\cdot 1}$, the weights appropriate to the z's, will be zero, and $\bar{z}_0$ will equal zero for every individual. In this case $k_{0\cdot 12} = \sigma_z = 1.0$.

This is the maximum value that k can ever take and means that the error of estimate has not been reduced at all by the use of z_1 and z_2 over what it would be were sheer random guesses resorted to. If z_0 can be perfectly estimated from z_1 and z_2 then every $(z_0 - \bar{z}_0)$ equals zero and $k_{0\cdot 12} = .00$. This is the minimum value that k can take and corresponds to perfect estimation, or zero errors of estimate throughout. In the symbol $k_{0\cdot 12}$ the subscript before the point designates the variable estimated and the subscripts after the point designate the variables from which the estimate has been made. The problem has been stated to determine the β's so that k shall be

a minimum. The constant $k_{0\cdot12}$ is the standard deviation of the errors of estimate when scores are expressed in terms of standard measures. Its meaning is thus easily grasped and obviously very important for the magnitude of the error involved in estimating one variable knowing all the others is the first item of information needed in interpreting the significance of the relation between variables. It will later be shown that $k_{0\cdot12}$ varies directly as $\sigma_{0\cdot12}$, the standard error of estimate of the x_0's, or the X_0's, so that establishing the minimal error condition with reference to the standard measures also establishes it with reference to the gross scores.

The following derivation of the values of the β's is brief and simple, but involves an understanding of calculus. For those unfamiliar with calculus a numerical illustration showing the concrete significance of the constants involved is given in the next section.

It is required to so choose $\beta_{01\cdot2}$ and $\beta_{02\cdot1}$ that the standard error of estimate shall be a minimum; that is,

$$\Sigma\,(z_0 - \bar{z}_0)^2 = \Sigma\,(z_0 - \beta_{01\cdot2}z_1 - \beta_{02\cdot1}z_2)^2$$

is to be a minimum. Differentiating first with respect to $\beta_{01\cdot2}$, and second with respect to $\beta_{02\cdot1}$, gives the two following equations

$$\Sigma\ 2\,(z_0 - \beta_{01\cdot2}z_1 - \beta_{02\cdot1}z_2)\,(-z_1) = 0$$
$$\Sigma\ 2\,(z_0 - \beta_{01\cdot2}z_1 - \beta_{02\cdot1}z_2)\,(-z_2) = 0$$

Dividing by $-2N$, summing the several parts, and remembering that

$$\frac{\Sigma\,z^2_1}{N} = \frac{\Sigma\,z^2_2}{N} = 1.0$$

that

$$\frac{\Sigma\,z_0z_1}{N} = r_{01}$$

that

$$\frac{\Sigma\,z_0z_2}{N} = r_{02}$$

and that

$$\frac{\Sigma\,z_1z_2}{N} = r_{12}$$

we obtain

$$\begin{aligned} r_{01} - \beta_{01\cdot2} - r_{12}\beta_{02\cdot1} &= 0 \\ r_{02} - r_{12}\beta_{01\cdot2} - \beta_{02\cdot1} &= 0 \end{aligned} \qquad \text{(Normal equations)} \ldots [241]$$

Solving simultaneously

$$\beta_{01\cdot 2} = \frac{r_{01} - r_{02}r_{12}}{1 - r^2_{12}}$$

$$\beta_{02\cdot 1} = \frac{r_{02} - r_{01}r_{12}}{1 - r^2_{12}}$$

(Regression coefficients between standard measures — 3 variables) .....[242]

This completes the solution of the 3-variable regression equation involving standard measures. We will make the usual transformations,

$$z = \frac{X - M}{\sigma}$$

and express the result in terms of gross scores, giving

$$\frac{\overline{X}_0 - M_0}{\sigma_0} = \beta_{01\cdot 2}\left(\frac{X_1 - M_1}{\sigma_1}\right) + \beta_{02\cdot 1}\left(\frac{X_2 - M_2}{\sigma_2}\right)$$

which, upon simplification, becomes,

$$\overline{X}_0 = \beta_{01\cdot 2}\frac{\sigma_0}{\sigma_1}X_1 + \beta_{02\cdot 1}\frac{\sigma_0}{\sigma_2}X_2 + \left(M_0 - \beta_{01\cdot 2}\frac{\sigma_0}{\sigma_1}M_1 - \beta_{02\cdot 1}\frac{\sigma_0}{\sigma_2}M_2\right). \quad [243]$$

Defining $b_{01\cdot 2}$, $b_{02\cdot 1}$, and c by the following equations

$$b_{01\cdot 2} = \beta_{01\cdot 2}\frac{\sigma_0}{\sigma_1}, \quad b_{02\cdot 1} = \beta_{02\cdot 1}\frac{\sigma_0}{\sigma_2} \quad .....[244]$$

$$c = M_0 - b_{01\cdot 2}M_1 - b_{02\cdot 1}M_2 \quad .....[245]$$

equation [243] may be written

$$\overline{X}_0 = b_{01\cdot 2}X_1 + b_{02\cdot 1}X_2 + c$$

(Regression equation involving gross scores — 3 variables) .[246]

Very simple algebraic derivation will show that in the case of n independent variables we have

$$b_{01\cdot 23\ldots n} = \beta_{01\cdot 23\ldots n}\frac{\sigma_0}{\sigma_1}$$

$$b_{02\cdot 13\ldots n} = \beta_{02\cdot 13\ldots n}\frac{\sigma_0}{\sigma_2} \quad .....[247]$$

in which $\beta_{01\cdot 23\ldots n}$, $\beta_{02\cdot 13\ldots n}$, etc., are defined by formula [264 b]

$$c = M_0 - b_{01\cdot 23\ldots n}M_1 - b_{02\cdot 13\ldots n}M_2 - \cdots - b_{0n\cdot 12\ldots n-1}M_n \quad .....[248]$$

$$\overline{X}_0 = b_{01\cdot 23\ldots n}X_1 + b_{02\cdot 13\ldots n}X_2 + \cdots b_{0n\cdot 12\ldots n-1}X_n + c \quad .....[249]$$

Equation [246] is ordinarily the most convenient form to use. The constants $b_{01\cdot 2}$, $b_{02\cdot 1}$ and c have numerical values which do not change for the entire population, and it only remains to substitute the gross scores, X_1 and X_2, to secure an estimate of the dependent gross score X_0.

We have determined the value of $\beta_{01\cdot 2}$ in terms of total correlation coefficients r_{01}, r_{02}, and r_{12}, and its use in the regression equation, but have still to discover the property which has led to the subscript notation. Let us find the regression of that part of z_0 which is independent of z_2 upon that part of z_1 which is independent of z_2. Since the regression equation connecting z_0 with z_2 is

$$\bar{z}_0 = r_{02}z_2$$

That part of z_0 which cannot be estimated from a knowledge of z_2, or that part which is independent of z_2, is $(z_0 - r_{02}z_2)$. This magnitude we will designate by $z_{0\cdot 2}$, which may be read "the residual in z_0 after estimation of z_0 by aid of z_2" or "that part of z_0 which is independent of z_2."

$$z_{0\cdot 2} = (z_0 - r_{02}z_2) \quad \text{(An error of estimate, i.e., a residual)} \quad \ldots\ldots\ldots [239\ a]$$

Obviously the N residuals, $z_{0\cdot 2}$ cannot be estimated at all by means of z_2, since z_2 has already been used for all that it avails. This is merely equivalent to saying that the regression of $z_{0\cdot 2}$ upon z_2 is equal to zero. The proof is simple:

$$b_{0\cdot 2,\,2} = \frac{\Sigma\, z_{0\cdot 2}z_2}{\Sigma\, z^2{}_2}$$

$$\Sigma\, z_{0\cdot 2}z_2 = \Sigma\,(z_0 - r_{02}z_2)\, z_2 = \Sigma\, z_0z_2 - r_{02}\,\Sigma\, z^2{}_2 = Nr_{02} - Nr_{02} = 0$$

accordingly $b_{0\cdot 2,\,2} = 0$. We may, however, estimate these residuals by means of variable 1 which is a new source of data. Since $z_{0\cdot 2}$ has zero regression upon z_2, it of course has zero regression upon that part of z_1 which can be estimated by means of z_2. To estimate z_1 from z_2 we have $\bar{z}_1 = r_{12}z_2$ so that

$$b_{(z_{0\cdot 2})\,(r_{12}z_2)} = \frac{\Sigma\,(z_{0\cdot 2})\,(r_{12}z_2)}{\Sigma\,(r_{12}z_2)^2} = \frac{r_{12}\,\Sigma\, z_{0\cdot 2}z_2}{\Sigma\,(r_{12}z_2)^2} = 0$$

It is therefore clear that only $z_{1\cdot 2}(= z_1 - r_{12}z_2)$, that part of z_1 which is independent of z_2, is of service in estimating $z_{0\cdot 2}$, that part of z_0 which is independent of z_2. The regression of $z_{0\cdot 1}$ upon $z_{1\cdot 2}$ is

$$\frac{\Sigma\, z_{0\cdot 2}z_{1\cdot 2}}{\Sigma\, z^2{}_{1\cdot 2}} = \frac{\Sigma\,(z_0 - r_{02}z_2)\,(z_1 - r_{12}z_2)}{\Sigma\,(z_1 - r_{12}z_2)^2}$$

$$= \frac{r_{01} - r_{02}r_{12} - r_{02}r_{12} + r_{02}r_{12}}{1 - 2\,r_{12}{}^2 + r_{12}{}^2}$$

$$= \frac{r_{01} - r_{02}r_{12}}{1 - r^2{}_{12}} = \beta_{01\cdot 2}$$

We now see the meaning of the notation $\beta_{01\cdot 2}$. It is the regression of that part of z_0 which is independent of z_2 upon that part of z_1 which is independent of z_2. For this reason $\beta_{01\cdot 2}$ is called a partial regression coefficient and, to recapitulate, it has the two following important properties:

(a) It is the regression of that part of z_0 which is independent of z_2 upon that part of z_1 which is independent of z_2.

(b) It is the weight or multiplying factor of z_1 when z_1 and z_2 are both used to estimate z_0.

Of course $\beta_{02\cdot 1}$ is the comparable partial regression coefficient when variables z_1 and z_2 are interchanged. We will now illustrate this by a numerical example.

Section 81. Three-variable Problem Illustrating Meanings of Constants

The first three columns of Table LX constitute the series to be correlated and the subsequent columns are derived calculations.

TABLE LX

z_0	z_1	z_2	$r_{02}z_2$	$z_{0\cdot 2}$	$r_{12}z_2$	$z_{1\cdot 2}$	$\beta_{01\cdot 2}z_{1\cdot 2}$	$z_{0\cdot 12}$	$\bar{z}_0$
1.75	1.00	.25	.1237	1.6263	−.0638	1.0638	.8667	.7596	.9904
1.25	.25	1.00	.4948	.7552	−.2552	.5052	.4116	.3436	.9064
1.00	.00	1.00	.4948	.5052	−.2552	.2552	.2079	.2973	.7027
.75	1.50	.00	.0000	.7500	.0000	1.5000	1.2221	−.4721	1.2221
.25	−.75	2.00	.9896	−.7396	−.5104	−.2396	−.1952	−.5444	.7944
.25	1.25	−.50	−.2474	.4974	.1276	1.1224	.9145	−.4171	.6671
−.25	.75	−1.25	.6185	.3685	.3190	.4310	.3512	.0173	−.2673
−.50	−1.00	.00	.0000	−.5000	.0000	−1.0000	−.8148	.3148	−.8148
−.75	.00	−1.00	−.4948	−.2552	.2552	−.2552	−.2079	−.0473	−.7027
−1.00	−1.00	.00	.0000	−1.0000	.0000	−1.0000	−.8148	−.1852	−.8148
−1.25	.00	−1.75	−.8659	−.3841	.4466	−.4466	−.3639	−.0202	−1.2298
−1.50	−2.00	.25	.1237	−1.6237	−.0638	−1.9362	−1.5775	−.0462	−1.4538

$\Sigma z_0 z_1 = 7.6250, \quad r_{01} = .63542$

$\Sigma z_0 z_2 = 5.9375, \quad r_{02} = .49479, \quad \beta_{01\cdot 2} = \dfrac{r_{01} - r_{02}r_{12}}{1 - r^2_{12}} = .81476$

$\Sigma z_1 z_2 = -3.0625, \quad r_{12} = -.25521, \quad \beta_{02\cdot 1} = \dfrac{r_{02} - r_{01}r_{12}}{1 - r^2_{12}} = .70272$

$\Sigma z_{0\cdot 2} z_{1\cdot 2} = 9.14030, \quad \bar{z}_0 = \beta_{01\cdot 2} z_1 + \beta_{02\cdot 1} z_2 = .81476\, z_1 + .70272\, z_2$

$\Sigma z^2_{1\cdot 2} = 11.21842$

$\beta_{01\cdot 2} = \dfrac{9.14030}{11.21842} = .81476 \qquad \sigma_{0\cdot 12} = \sqrt{\dfrac{1.61504}{12}} = .36686$

$\Sigma z^2_{0\cdot 12} = 1.61504$

These series have been so chosen that the means equal zero and the standard deviations equal one. We are thus dealing with standard measures, or z's and not with x's or X's. Straightforward calculation gives

$$r_{01} = \frac{\Sigma z_0 z_1}{N \text{ 1.0} \times \text{1.0}} = \frac{7.625}{12} = .63542$$

$$r_{02} = .49479$$

$$r_{12} = -.25521$$

We can estimate z_0 by means of z_2 by the following equation:

$$\bar{z}_0 = r_{02} z_2 = .49479\, z_2$$

These estimated values are recorded in the column $r_{02}z_2$. The residuals $(z_0 - r_{02}z_2)$, or parts of z_0 which are independent of z_2, are recorded in the column $z_{0.2}$. We can estimate z_1 by means of z_2 by the equation

$$\bar{z}_1 = r_{12} z_2 = -.25521\, z_2$$

These estimated values are recorded in column $r_{12}z_2$. The residuals $(z_1 - r_{12}z_2)$, or parts of z_1 independent of z_2, are recorded in the column $z_{1.2}$. That part of z_1 which is independent of z_2, namely $z_{1.2}$, may be used to estimate $z_{0.2}$. Straightforward calculation of the regression equation gives

$$\bar{z}_{0.2} = \frac{\Sigma z_{0.2} z_{1.2}}{\Sigma z^2_{1.2}} z_{1.2} = \frac{9.14030}{11.21842} z_{1.2} = .81476\, z_{1.2}$$

The constant .81476 $(= \beta_{01.2})$ is here seen to be a regression coefficient, being just as real and definite in its meaning as those found in any other two-variable problem. Finally taking $(z_{0.2} - \beta_{01.2}z_{1.2})$ we obtain $z_{0.12}$, the final residuals that are left after having utilized both z_1 and z_2 to the utmost in estimating z_0. These magnitudes are our final errors of estimate. Calculating their standard deviation in the usual manner we obtain

$$k_{0.12} = .36686$$

The residuals $z_{0.12}$ could have been obtained more directly without the calculation of $z_{0.2}$ and $z_{1.2}$ by the regression equation involving the two variables. We have

$$z_{0.12} = z_0 - \beta_{01.2}z_1 - \beta_{02.1}z_2$$

in which

$$\beta_{01\cdot 2} = \frac{r_{01} - r_{02}r_{12}}{1 - r^2_{12}} = .81476$$

$$\beta_{02\cdot 1} = \frac{r_{02} - r_{01}r_{12}}{1 - r^2_{12}} = .70272$$

The more lengthy procedure has been followed for the purpose of showing the exact significance of the β constants and of the residuals, and not because it is the most practical method for purposes of estimation. If we add the measures in the two columns $r_{02}z_2$ and $\beta_{01\cdot 2}z_{1\cdot 2}$ or if we use equation

$$\bar{z}_0 = \beta_{01\cdot 2}z_1 + \beta_{02\cdot 1}z_2$$

we obtain the best estimates of z_0 which it is possible to secure from z_1 and z_2, assuming a rectilinear relationship. Such estimates are here recorded in column $\bar{z}_0$. The correlation between z_0 and $\bar{z}_0$ is the multiple correlation coefficient and will be designated by the symbol $r_{0\cdot 12}$. As multiplying every term in a series by a constant, or adding a constant amount to every term, does not change the correlation with a second variable, the correlation between z_0 and $\bar{z}_0$ is identical with that between x_0 and $\bar{x}_0$ or between X_0 and $\bar{X}_0$. The multiple correlation coefficient is the maximum correlation obtainable between dependent variable and a weighted composite of the independent variables. We may therefore read $r_{0\cdot 12}$ as "the correlation between the variable 0 and the best weighted linear combination of variables 1 and 2." Straightforward calculation of the correlation between columns z_0 and $\bar{z}_0$ yields $r_{0\cdot 12} = .93028$, but a much shorter method of calculation is available. We have in a two-variable problem

$$\sigma_{1\cdot 2} = \sigma_1\sqrt{1 - r^2}$$

Since z_0 and $\bar{z}_0$ are simply two variables and since the standard deviation of $z_0 = 1.0$, and the standard deviation of the residuals in z_0 after estimation by aid of z_1 and z_2 is $k_{0\cdot 12}$ we have

$$k_{0\cdot 12} = 1.0\sqrt{1 - r^2_{0\cdot 12}}$$

from which

$$r_{0\cdot 12} = \sqrt{1 - k^2_{0\cdot 12}}$$ (Value of the multiple correlation coefficient — 3 variables) . . . .[250]

The relation between $k_{0\cdot 12}$ and $r_{0\cdot 12}$ is the same as that between k_{12} and r_{12} of Formula [86 *a*], section 48, hence $k_{0\cdot 12}$ is a coeffi-

cient of alienation in the case of three variables. We now need a simple procedure for the calculation of $k_{0.12}$. Since $k_{0.12}$ is the standard deviation of the residuals we have

$$k^2_{0.12} = \frac{1}{N}\Sigma\,(z_0 - \beta_{01.2}z_1 - \beta_{02.1}z_2)^2$$

Squaring, summing, and collecting terms we will find that the factor $(1 - r^2_{12})$ enters into numerator and denominator. Wherever this factor occurs we will write k^2_{12}. Remembering that

$$\Sigma\, x^2_0 = \Sigma\, z^2_1 = \Sigma\, z^2_2 = N$$

and that

$$\Sigma\, z_0 z_1 = N r_{01},\ \Sigma\, z_0 z_2 = N r_{02},\ \Sigma\, z_1 z_2 = N r_{12}$$

we have

$$\begin{aligned} k^2_{0.12} &= 1 + \beta^2_{01.2} + \beta^2_{02.1} - 2\,\beta_{01.2}r_{01} - 2\,\beta_{02.1}r_{02} \\ &\quad + 2\,\beta_{01.2}\beta_{02.1}r_{12} \\ &= \frac{1}{k^2_{12}}\,(1 - r^2_{01} - r^2_{02} - r^2_{12} + 2\,r_{01}r_{02}r_{12}) \end{aligned}$$

(Coefficient of alienation — 3 variables) . . [251]

The general solution of the coefficient of alienation in the case of n variables is well accomplished by the aid of determinants, and we may here note this form of solution for the case of three variables. If we write the major determinant

$$\Delta = \begin{vmatrix} 1 & r_{01} & r_{02} \\ r_{01} & 1 & r_{12} \\ r_{02} & r_{12} & 1 \end{vmatrix} \quad \ldots\ldots\ldots\ldots\ldots\ldots [252]$$

and call the minor obtained by deleting the first row and the first column Δ_{00}, we have

$$\Delta_{00} = \begin{vmatrix} 1 & r_{12} \\ r_{12} & 1 \end{vmatrix} = k^2_{12} \quad \ldots\ldots\ldots\ldots\ldots\ldots [253]$$

Evaluating these determinants we obtain the numerator and denominator respectively of the fraction giving $k^2_{0.12}$ so that we may write

$$k_{0.12} = \sqrt{\frac{\Delta}{\Delta_{00}}} \qquad \begin{array}{l}\text{(Multiple coefficient of alienation as} \\ \text{the quotient of determinants)}\end{array} \ldots [254]$$

This is here proven for the case of three variables, but we will later find that the equation holds generally for any number of variables. If we are concerned only with the value of the multiple correlation coefficient, and not with the constants of the regression equation, the simplest way to find it is to first

determine $k_{0\cdot12}$ and then $r_{0\cdot12}$. If we have the regression coefficients we may obtain $k_{0\cdot12}$ and thus $r_{0\cdot12}$ from it. We have called $k_{0\cdot12}$ the multiple alienation coefficient. It is the measure of independence of variable x_0 from variables x_1 and x_2. We will define $k_{01\cdot2}$ as the partial alienation coefficient. It is the measure of independence of x_0 and x_1 for a constant value of x_2. Thus, by definition, if $r_{01\cdot2}$ is the partial correlation between x_0 and x_1 for a constant value of z_2, we have

$$k^2_{01\cdot2} + r^2_{01\cdot2} = 1.0$$ (Relation between partial coefficients of correlation and of alienation) . . . [255]

This is the equation for three variables comparable to formula [86 *a*], $k^2_{12} + r^2_{12} = 1.0$, found for two variables. We thus find that whether k has one primary subscript (a subscript occurring before the point is termed a primary and one after the point a secondary subscript), $k_{0\cdot12}$, or too primary subscripts, $k_{01\cdot2}$ the type equation, $k^2 + r^2 = 1.0$, holds. Thus far we have found the total, multiple, and partial relationships as follows, respectively.

$$k^2_{01} + r^2_{01} = 1$$
$$k^2_{0\cdot12} + r^2_{0\cdot12} = 1$$
$$k^2_{01\cdot2} + r^2_{01\cdot2} = 1$$

The same relation will be found to hold when n variables are involved, so that universally, provided the subscripts are the same,

$$k^2 + r^2 = 1$$ (General relation between k and r) . . . . . [256]

We do not have a k with three primary subscripts, but k_{01} and $k_{0\cdot1}$ may be shown to be identical. Dealing with z's we have found $k_{01} = \sqrt{1 - r^2_{01}}$ and $k_{0\cdot1}$ = the standard deviation of the arrays of z_0's, i.e., $k_{0\cdot1} = \sigma_{z_0}\sqrt{1 - r^2_{01}} = \sqrt{1 - r^2_{01}}$, since, when dealing with z's the standard deviation σ_z is equal to 1. Accordingly

$$k_{0\cdot1} = k_{01}. \quad \ldots\ldots\ldots\ldots\ldots\ldots [257]$$

Equations [251] and [254] have expressed $k_{0\cdot12}$ in terms of the total correlation coefficients. We may also evaluate this multiple alienation coefficient in terms of other total and partial coefficients, but will first need to determine a partial coefficient of correlation. Having shown that $\beta_{01\cdot2}$ is the regression of

$z_{0\cdot2}$ upon $z_{1\cdot2}$ and since by parity $\beta_{10\cdot2}$ is the regression of $z_{1\cdot2}$ upon $z_{0\cdot2}$, we immediately have, since every condition leading to $r_{12} = \sqrt{b_{12}b_{21}}$ formula [90] is exactly paralleled when dealing with $z_{0\cdot2}$'s and $z_{1\cdot2}$'s,

$$r_{01\cdot2} = \sqrt{\beta_{01\cdot2}\beta_{10\cdot2}}$$

(Partial coefficient of correlation in terms of partial regression coefficients — 3 variables) .....[258]

The partial coefficient $r_{01\cdot2}$ is identical with $r_{10\cdot2}$ but custom places first the numerically smaller of the subscripts before the point.

$$k^2_{01\cdot2} = 1 - \beta_{01\cdot2}\beta_{10\cdot2} = \frac{1 - r^2_{01} - r^2_{02} - r^2_{12} + 2\,r_{01}r_{02}r_{12}}{k^2_{12}k^2_{02}}$$

$$k^2_{02}k^2_{01\cdot2} = \frac{1 - r^2_{01} - r^2_{02} - r^2_{12} + 2\,r_{01}r_{02}r_{12}}{k^2_{12}} = k^2_{0\cdot12}$$

that is

$$k_{0\cdot12} = k_{02}k_{01\cdot2}$$

or

$$k_{0\cdot12} = k_{01}k_{02\cdot1}$$

(Multiple coefficient of alienation in terms of total and partial coefficients of alienation — 3 variables) .....[259]

We may now outline the most expeditious manner of calculating all of the constants ordinarily desired in the solution of a multiple correlation problem. These constants recorded in the proper order of calculation are:

the means, M_0, M_1, and M_2

the standard deviations, σ_0, σ_1, and σ_2

the total correlations, r_{01}, r_{02} and r_{12}

the squares of the total alienation coefficients k^2_{01}, k^2_{02} and k^2_{12}

the β regression coefficients

$$\beta_{01\cdot2} = \frac{r_{01} - r_{02}r_{12}}{k^2_{12}}, \quad \beta_{10\cdot2} = \frac{r_{01} - r_{02}r_{12}}{k^2_{02}}, \quad \beta_{02\cdot1} = \frac{r_{02} - r_{01}r_{12}}{k^2_{12}}$$

the square of the partial correlation coefficient

$$r^2_{01\cdot2} = \beta_{01\cdot2}\,\beta_{10\cdot2}$$

the square of the partial alienation coefficient

$$k^2_{01\cdot2} = 1 - r^2_{01\cdot2}$$

the square of the multiple alienation coefficient

$$k^2_{0\cdot12} = k^2_{02}k^2_{01\cdot2}$$

the multiple alienation coefficient, $k_{0\cdot12}$

the multiple correlation coefficient, $r_{0\cdot12} = \sqrt{1 - k^2_{0\cdot12}}$

the b regression coefficients

$$b_{01\cdot 2} = \beta_{01\cdot 2}\frac{\sigma_0}{\sigma_1},\quad b_{02\cdot 1} = \beta_{02\cdot 1}\frac{\sigma_0}{\sigma_2}$$

the constant c

$$c = M_0 - b_{01\cdot 2}M_1 - b_{02\cdot 1}M_2$$

giving the regression equation

$$\overline{X}_0 = b_{01\cdot 2}X_1 + b_{02\cdot 1}X_2 + c$$

the standard error of estimate, or the standard deviation of the X_0-arrays from the regression line

$$\sigma_{0\cdot 12} = \sigma_0 k_{0\cdot 12}$$

Excepting the probable errors of the constants (see formulas [278], [279] and [280]) the solution is complete.

Section 82. The Use of the Alignment Chart

The calculation of the β constants may be easily accomplished by the aid of an alignment chart. The following directions apply to the small chart in the appendix and described in detail and with explanatory problems in (Kelley, 1921, chart), and also to a large chart devised upon the same principle (Kelley, 1921, align). Items (*i*) and (*j*) and the four-variable problem illustration should be read after the treatment of the n variable problem, Section 83, of this text. The accuracy of the chart in the appendix is very slightly less than that of a 10-inch slide rule, while the large chart gives results of the same degree of accuracy as a 20-inch slide rule.

The scales for r_{13} and r_{23} are graduated according to the logarithms of numbers from 10 to 100, and the product scale is so graduated as to indicate the products of any two numbers on scales r_{13} and r_{23} when connected by a straight line. Accordingly all products and quotients, including squares and square roots, may be obtained. In all these operations the simplest way to keep track of the decimal point is to roughly carry the operation through in one's head and then place the point where it belongs. A strip of transparent celluloid with a straight line scratched upon it, or a silk thread drawn taut, constitute serviceable straight edges.

Scale $1/k$ is graduated according to the logarithms of $1/\sqrt{1-r^2}$ and scale $1/k^2$ according to the logarithms of $1/1-r^2$. Scale $1/K^2$ is a continuation of scale $1/k^2$. When

values on scale $1/K^2$ are used, place a straight edge through this value and parallel to the base line [as explained in example (*c*)] and locate a point on scale $1/k^2$. Then continue the calculation using the point so located on scale $1/k^2$ in lieu of the point on scale $1/K^2$.

The following magnitudes are needed in multiple correlation work:

(*a*) Products, such as $r_{13}r_{23}$

(*b*) Quotients, such as $\frac{\sigma_1}{\sigma_2}$

(*c*) Square roots, such as $\sqrt{\beta_{12\cdot3}\beta_{21\cdot3}}$

(*d*) Factors $\frac{1}{k_{13}}\left(=\frac{1}{\sqrt{1-r^2_{13}}}\right)$ which enter into partial coefficients of correlation

(*e*) Coefficients of alienation, such as $k_{13}\ (=\sqrt{1-r^2_{13}})$

(*f*) Factors $\frac{1}{k^2_{23}}\left(=\frac{1}{1-r^2_{23}}\right)$ which enter into regression coefficients

(*g*) Squares of coefficients of alienation, such as k^2_{23} $(=1-r^2_{23})$

(*h*) Partial regression coefficients, such as

$$\beta_{12\cdot3}\left(=\frac{r_{12}-r_{13}r_{23}}{k^2_{23}}\right) \qquad [247]$$

(*i*) Partial correlation coefficients, such as

$$r_{12\cdot3}\left(=\frac{r_{12}-r_{13}r_{23}}{k_{13}k_{23}}=\sqrt{\beta_{12\cdot3}\beta_{21\cdot3}}=\sqrt{b_{12\cdot3}b_{21\cdot3}}\right)$$

(*j*) Partial regression coefficients involving four variables

$$\beta_{12\cdot34}\left(=\frac{\beta_{12\cdot4}-\beta_{13\cdot4}\beta_{32\cdot4}}{k^2_{23\cdot4}}=\frac{\beta_{12\cdot3}-\beta_{14\cdot3}\beta_{42\cdot3}}{k^2_{24\cdot3}}\right)$$

Since $k^2_{23\cdot4}=1-\beta_{23\cdot4}\beta_{32\cdot4}$, and since the calculation which leads to $\beta_{23\cdot4}$ is changed in but one simple respect to obtain $\beta_{32\cdot4}$ it is convenient to write:

$$\beta_{12\cdot34}=\frac{\beta_{12\cdot4}-\beta_{13\cdot4}\beta_{32\cdot4}}{1-\beta_{23\cdot4}\beta_{32\cdot4}} \qquad [264\ a]$$

(*k*) Partial regression coefficients involving more than four variables

$$\beta_{12\cdot34\ldots n}=\frac{\beta_{12\cdot4\ldots n}-\beta_{13\cdot4\ldots n}\beta_{32\cdot4\ldots n}}{1-\beta_{23\cdot4\ldots n}\beta_{32\cdot4\ldots n}} \qquad [264\ b]$$

The same procedure as in (*j*) is followed, but in this

case the calculation which leads to $\beta_{23.4\ldots n}$ does not, by one simple change, lead to $\beta_{32.4\ldots n}$.

Examples:

(*a*) $.2 \times .4$ Place a straight edge on 20, scale r_{13}, and upon 40, scale r_{23}, and read the product, .08, on the product scale.

(*b*) $\frac{2}{.4}$ Place a straight edge upon 20, product scale, and upon 40, scale r_{23}, and read the quotient, 5.0, on scale r_{13}.

(*c*) $\sqrt{.25}$ Place a straight edge on 25, product scale, and parallel to the base line of the chart (this can be done by rotating the straight edge until the readings on scales r_{13} and r_{23} are identical) and read the square root, .50, on either scale r_{13} or r_{23}.

(*d*) $\frac{1}{\sqrt{1 - .60^2}}$ Find 60 on scale $1/k$ and read the answer, 1.25, from the same point on scale r_{13}.

(*e*) $\sqrt{1 - .60^2}$ Place a straight edge through 60, scale $1/k$, and 100, product scale, and read the answer, .80, on scale r_{23}.

(*f*) $\frac{1}{1 - .60^2}$ Find 60 on scale $1/k^2$ and read the answer, 1.5625, from the same point on scale r_{23}.

(*g*) $1 - .60^2$ Place a straight edge through 60, scale $1/k^2$, and 100, product scale, and read the answer, .64, on scale r_{13}.

(*h*) $\frac{.78 - .60 \times .80}{1 - 80^2}$ Find the product of .60 and .80 by (*a*). On a separate scratch paper subtract this from .78, obtaining .30. Place a straight edge between 30, scale r_{13}, and 80, scale $1/k^2$, and read the answer, .833, on the product scale.

(*i*) $\frac{.78 - .60 \times .80}{\sqrt{1 - 60^2}\sqrt{1 - .80^2}}$ Find $\frac{.78 - .60 \times .80}{1 - .80^2}$ by (*h*).

Find $\frac{.78 - .60 \times .80}{1 - .60^2}$ by (*h*). Multiply and extract the square root by (*a*) and (*c*), yielding the answer .625.

(*j*) Given: $\beta_{12.4} = .70$; $\beta_{13.4} = 60$; $\beta_{32.4} = .80$; $\beta_{23.4} = .5469$.

Required: $\beta_{12.34} = \dfrac{.70 - .60 \times .80}{1 - .80 \times .5469}$ Find the numerator as in (*h*) and the denominator in the same manner. Then divide as in (*b*). This gives $\dfrac{.2200}{.5625} = .3911$.

If, as is frequently the case, $\beta_{32.4}$ and $\beta_{23.4}$ are nearly equal, $k^2_{23.4}$ is closely given by:

$$k^2_{23.4} = 1 - \left(\frac{\beta_{32.4} + \beta_{23.4}{}^2}{2}\right)$$

In this case the procedure may be as follows:

$$\frac{.70 - .60 \times .80}{1 - .80 \times .76}$$

Find the numerator, .2200, as before. On scratch paper determine .78, the arithmetic average of .80 and .76. Place a straight edge between .78, scale $1/k^2$, and .22, scale r_{13}, and read the answer, .5618, on the product scale. This answer is in error by .0006, which is of the same order of magnitude as the error attendant upon the use of the large chart.

As a sample problem in three variables the following data are given:

TABLE LXI

Table of Correlations, Means and Standard Deviations

	Variables		
	1	2	3
2	.225		
3	.274	.404	
Means	68.15	43.60	52.20
σ's	10.50	12.24	9.63

Solving

$$\beta_{12.3} = .1366$$
$$\beta_{21.3} = .1236$$
$$\beta_{13.2} = .2200$$
$$k^2_{1.23} = .9093$$
$$r_{1.23} = .3011$$
$$\sigma_{1.23} = 10.01$$

$$\bar{z}_1 = .1366\, z_2 + .2200\, z_3$$

$$\overline{X_1} = .1172\, X_2 + .2399\, X_3 + 50.52$$

As a sample problem in four variables the following data are given:

TABLE LXII

	Variables			
	1	2	3	4
2	.225			
3	.274	.404		
4	.134	.060	.231	
Means	68.15	43.60	52.20	45.40
σ's	10.50	12.24	9.63	14.25

Solving

$$\beta_{12\,34} = .1398$$
$$\beta_{21\,34} = .1270$$
$$\beta_{13\,24} = .1991$$
$$\beta_{14\,23} = .0796$$
$$k^2_{1\,234} = .9033$$
$$r_{1\,234} = .3109$$
$$\sigma_{1\,234} = 9.980$$

$$\bar{z}_1 = .1398\, z_2 + .1991\, z_3 + .0796\, z_4$$
$$\overline{X_1} = .1199\, X_2 + .2171\, X_3 + .0587\, X_4 + 48.92$$

Section 83. The General Treatment of the n-Variable Problem

We will now attack the general problem. The reader will need an elementary knowledge of determinants to follow the discussion. We are given a criterion variable, X_0, and the independent variables, $X_1, X_2, \cdots, X_n$ (the population will be designated by N, which symbol must not be confused with n, the number of independent variables). Expressing every variable in terms of standard measures by the transformations

$$z = \frac{X - M}{\sigma}$$

it is required to determine the β constants in the following equation in the best fit manner.

$$\bar{z}_0 = \beta_{01\cdot 23\ldots n}\, z_1 + \beta_{02\cdot 13\ldots n}\, z_2 + \cdots + \beta_{0n\cdot 12\ldots n-1}\, z_n \quad \ldots[260]$$

$(z_0 - \bar{z}_0)$ is an error of estimate and will be designed by $z_{0\cdot 12\ldots n}$.

The β's are to be so determined that the standard error of these errors of estimate, $k_{0\cdot 12\ldots n}$, shall be a minimum.

$$k^2_{0\cdot 12\ldots n} = \frac{1}{N}\Sigma\, z^2{}_{0\ 12\ldots n}$$

$$= \frac{1}{N}\Sigma\,(z_0 - \beta_{01\cdot 23\ldots n}\, z_1 - \beta_{02\ 13\ldots n}\, z_2 - \cdots - \beta_{0n\cdot 12\ldots n-1})^2$$

Differentiating with respect to the first β and setting the derivative equal to zero, gives

$$\frac{2}{N}\Sigma\,[z_0 - \beta_{01\cdot 23\ldots n}\, z_1 - \beta_{02\cdot 13\ldots n}\, z_2 - \cdots - \beta_{0n\cdot 12\ldots n-1}\, z_n]\,(-z_1) = 0$$

Summing, expressing square sums in terms of standard deviations and product sums in terms of correlations, yields,

$$r_{01} - \beta_{01\ 23\ldots n} - r_{12}\beta_{02\ 13\ldots n} - \cdots - r_{1n}\beta_{0n\cdot 12\ldots n-1} = 0$$

Differentiating successively with respect to the other β's gives

$$r_{02} - r_{12}\beta_{01\ 23\ldots n} - \beta_{02\ 13\ldots n} - \cdots - r_{2n}\beta_{0n\cdot 12\ldots n-1} = 0$$

etc. to

$$r_{0n} - r_{1n}\beta_{01\ 23\ldots n} - r_{2n}\beta_{02\ 13\ldots n} - \cdots - \beta_{0n\cdot 12\ldots n-1} = 0$$

(Normal equations)...[261]

This gives n linear equations from which to determine the same number of β constants. The determinantal solution is readily written. Let the major determinant be Δ.

$$\Delta = \begin{vmatrix} 1 & r_{01} & r_{02} & \cdots & r_{0n} \\ r_{01} & 1 & r_{12} & \cdots & r_{1n} \\ r_{02} & r_{12} & 1 & \cdots & r_{2n} \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ r_{0n} & r_{1n} & r_{2n} & \cdots & 1 \end{vmatrix} \quad \ldots\ldots\ldots\ldots\ldots[262]$$

and let Δ_{pq} be the minor obtained by crossing out the p'th row and q'th column of the major determinant. The p'th row is that row having p as one of the subscripts of the r's throughout and the qth column is that column having q as one of the subscripts throughout. Then

$$\beta_{0p\cdot 12\ldots()\ldots n} = \frac{-(-1)^p\,\Delta_{0p}}{\Delta_{00}}$$ (The regression coefficient as the quotient of two determinants)..[263]

The quantity $-(-1)^p$ is merely a sign factor. The column

crossed out is the o'th for all the β's so that $q = 0$. To illustrate in detail we have

$$\beta_{01\cdot 23\cdots n} = \frac{\Delta_{01}}{\Delta_{00}} = \frac{\begin{vmatrix} r_{01} & r_{12} & r_{13} & \cdots & r_{1n} \\ r_{02} & 1 & r_{23} & \cdots & r_{2n} \\ r_{03} & r_{23} & 1 & \cdots & r_{3n} \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ r_{0n} & r_{2n} & r_{3n} & \cdots & 1 \end{vmatrix}}{\begin{vmatrix} 1 & r_{12} & r_{13} & \cdots & r_{1n} \\ r_{12} & 1 & r_{23} & \cdots & r_{2n} \\ r_{13} & r_{23} & 1 & \cdots & r_{3n} \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ r_{1n} & r_{2n} & r_{3n} & \cdots & 1 \end{vmatrix}}$$

$$\beta_{02\cdot 13\cdots n} = \frac{-\Delta_{02}}{\Delta_{00}} = \frac{\begin{vmatrix} r_{01} & 1 & r_{13} & r_{14} & \cdots & r_{1n} \\ r_{02} & r_{12} & r_{23} & r_{24} & \cdots & r_{2n} \\ r_{03} & 13 & 1 & r_{34} & \cdots & r_{3n} \\ \cdot & \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & \cdot & & \cdot \\ r_{0n} & r_{1n} & r_{3n} & r_{4n} & \cdots & 1 \end{vmatrix}}{\Delta_{00}} \qquad [264]$$

$$\beta_{03\cdot 124\cdots n} = \frac{\Delta_{03}}{\Delta_{00}}$$

$$\beta_{04\cdot 1235\cdots n} = \frac{-\Delta_{04}}{\Delta_{00}}$$

etc. to

$$\beta_{0n\cdot 12\cdots n-1} = \frac{-(-1)^n \Delta_{0n}}{\Delta_{00}}$$

Algebraic manipulation (see Kelley, 1921, chart) enables the expressing of a partial regression coefficient in terms of partial regression coefficients of one lower order, thus,

$$\beta_{12\cdot 34} = \frac{\beta_{12\cdot 4} - \beta_{13\cdot 4}\beta_{32\cdot 4}}{1 - \beta_{23\cdot 4}\beta_{32\cdot 4}} = \frac{\beta_{12\cdot 3} - \beta_{14\cdot 3}\beta_{42\cdot 3}}{1 - \beta_{24\cdot 3}\beta_{42\cdot 3}} \qquad \ldots\ldots [264\,a]$$

and in general

$$\beta_{12\cdot 34\cdots n} = \frac{\beta_{12\cdot 4\cdots n} - \beta_{13\cdot 4\cdots n}\beta_{32\cdot 4\cdots n}}{1 - \beta_{23\cdot 4\cdots n}\beta_{32\cdot 4\cdots n}} \qquad \ldots\ldots\ldots\ldots [264\,b]$$

Note that if the variables are designated by subscripts 1, 2, 3, $\cdots$ instead of as here, by 0, 1, 2$\cdots$ the sign factor is given

by $-(-1)^{p+q}$ in which q always equals 1. Probably the simplest way to keep track of the sign is to note that the denominator determinant is always positive and that the numerator determinants alternate in sign beginning with plus for the first β. Let us define

and

$$\left.\begin{matrix}\beta_{pq\cdot 12\ldots()\ldots()\ldots n}\\ \beta_{qp\cdot 12\ldots()\ldots()\ldots n}\end{matrix}\right\} \text{(Conjugate } \beta\text{'s)} \ldots . [265]$$

as conjugate regression coefficients. Then

$$\beta_{pq\cdot 12\ldots()\ldots()\ldots n} = \frac{-(-1)^{p+q}\Delta_{pq}}{\Delta_{pp}}$$

and

$$\beta_{qp\cdot 12\ldots()\ldots()\ldots n} = \frac{-(-1)^{q+p}\Delta_{qp}}{\Delta_{qq}}$$

Since the major determinant is symmetrical $\Delta_{pq} = \Delta_{qp}$ and the signs of the two are alike; thus the partial correlation coefficient is given by the square root of the product.

$$r_{pq\cdot 12\ldots()\ldots()\ldots n} = \frac{-(-1)^{p+q}\Delta_{pq}}{\sqrt{\Delta_{pp}}\sqrt{\Delta_{qq}}}$$ (Determinantal expression for the partial coefficient of correlation) . . . . . [266]

The partial correlations that are of most interest and value are generally those involving the criterion and required in the calculation of the multiple alienation coefficient.

$$r_{01\cdot 23\ldots n} = \frac{\Delta_{01}}{\sqrt{\Delta_{00}}\sqrt{\Delta_{11}}}$$ (A partial correlation coefficient of the $(n-1)$th order) . . . . . [267]

This may be written (Kelley, 1921, chart)

$$r_{01\cdot 23\ldots n} = \sqrt{\beta_{01\cdot 23\ldots n}\,\beta_{10\cdot 23\ldots n}} \ldots . [267\,a]$$

The order is determined by the number of secondary subscripts, thus $r_{01\cdot 2345}$ is a partial coefficient of the 4th order, $r_{01\cdot 2}$ of the first order and r_{01} of the zero order.

$$r_{02\cdot 34\ldots n} = \frac{\Delta_{01,\,12}}{\sqrt{\Delta_{01,\,01}}\sqrt{\Delta_{12,\,12}}}$$ (Determinantal expression for a partial correlation coefficient of the $n-2$ order) . . . [268]

The magnitude $\Delta_{01,\,12}$ indicates the minor obtained by crossing out the 0 and 1 row and the 1 and 2 column. Note that the sign factor is positive. This is clearly the case, since we are now really dealing with a major determinant of an order one lower in which row and column 2 have taken the place of row and column 1, row and column 3 the place of row and column 2, etc.

Continuing

$$r_{03\cdot 45\ldots n} = \frac{\Delta_{012,\ 123}}{\sqrt{\Delta_{012,\ 012}}\sqrt{\Delta_{123,\ 123}}}$$

etc. to

$$r_{0n} = \frac{\Delta_{012\ldots n-1,\ 123\ldots n}}{\sqrt{\Delta_{012\ldots n-1,\ 012\ldots n-1}}\sqrt{\Delta_{123\ldots n,\ 123\ldots n}}} = \frac{r_{0n}}{1 \times 1}$$

(Partial coefficient of zero order, or a total correlation coefficient) . . [269]

The various minors needed in the solution of this series of partial coefficients of correlation may be obtained incidentally in the process of obtaining the first minor if the determinant is evaluated in a certain manner which, however, may not always be the most convenient way for other needs. Having the various partial correlation coefficients we may determine the partial alienation coefficients by the equation $k = \sqrt{1 - r^2}$. These will prove serviceable in obtaining the multiple correlation coefficient, but we shall first need to establish the value of an alienation coefficient of a certain order in terms of an order one less. In dealing with z_0 and z_1 between which the correlation is r_{01} we have found, formula [257]

$$k^2_{0\cdot 1} = \sigma^2 z_0\,(1 - r^2_{01}) = 1\,(1 - r^2_{01}) = k^2_{01}$$

If we deal with magnitudes $z_{0\cdot 2}$, residuals in z_0, after estimation by z_2, and $z_{1\cdot 2}$ residuals in z_1 after estimation by z_2 between which the correlation is $r_{01\cdot 2}$ we have, following the identical reasoning that led to the preceding equation,

$$k^2_{0\cdot 12} = k^2_{0\cdot 2}\,(1 - r^2_{01\cdot 2}) = k^2_{0\cdot 2}\,k^2_{01\cdot 2} = k^2_{02}\,k^2_{01\cdot 2}$$

Obviously the principle can be applied to residuals of any order so that, in general,

$$k^2_{0\cdot 12\ldots n} = k^2_{0\cdot 23\ldots n}\,k^2_{01\cdot 23\ldots n}$$
$$k^2_{0\cdot 12\ldots n} = k^2_{0\cdot 13\ldots n}\,k^2_{02\cdot 13\ldots n}$$

etc. to

$$k^2_{0\cdot 12\ldots n} = k^2_{0\cdot 12\ldots\ n-1}\,k^2_{0n\cdot 12\ldots n-1}$$

(The n ways of expressing a multiple alienation coefficient of the n-th order, in terms of multiple alienation coefficients of the $(n\text{-}1)$th order and of partial alienation coefficients of the $(n\text{-}1)$th order) . . . . . [270]

Expressing $k^2_{0\cdot 23\ \ldots\ n}$ as equal to $k^2_{0\cdot 34\ \ldots\ n}\,k^2_{02\cdot 34\ \ldots\ n}$ and continuing the process for every k, until finally $k^2_{0\cdot n} = k^2_{0n}$, we have, taking the square root,

$$k_{0\cdot 12\ldots n} = k_{01\cdot 23\ldots n}\,k_{02\cdot 34\ldots n}\,k_{03\cdot 45\ldots n} \times \ldots \times k_{0n}$$

(One of the many ways of expressing a multiple alienation coefficient of the n-th order in terms of partial alienation coefficients of lower order) . . . . . [271]

Having the multiple alienation coefficient we obtain

$$r_{0\cdot 12\ldots n} = \sqrt{1 - k^2_{0\cdot 12\ldots n}} \quad \text{(The multiple correlation coefficient)}..[272]$$

and also

$$\sigma_{0\cdot 12\ldots n} = \sigma_0 k_{0\cdot 12\ldots n} \quad \text{(Standard error of estimate)}.. \ldots\ldots[273]$$

This completes the solution, but it is sometimes easier to obtain $r_{0\cdot 12\ldots n}$ by the direct evaluation of the major determinant Δ and the minor Δ_{00}. That we can obtain the multiple correlation coefficient in this manner will now be shown. If z_0 is the criterion and $\bar{z}_0$ the estimate of it, the correlation between them is the multiple correlation coefficient, and, if we let σ_- represent the standard deviation of the $\bar{z}_0$ measures, it is given by

$$r_{0\cdot 12\ldots n} = \frac{\Sigma z_0 \bar{z}_0}{N\sigma_0\sigma_-}$$

The standard deviation of the z_0 measures is the standard deviation of the points upon the regression line passing as closely as possible to the z_0 measures. Thus, just as in the case of two variables where $\sigma^2_1 = \sigma^2_{1\cdot 2} + \sigma^2_a$ [formula 87] in which σ_a is the standard deviation of the means of the arrays, so here with $(n + 1)$ variables.

$$\sigma^2_0 = \sigma^2_{0\cdot 12\ldots n} + \sigma^2_-$$

Dealing with z measures $\sigma_0 = 1$ and $\sigma_{0\cdot 12\ldots n} = k_{0\cdot 12\ldots n}$, so that

$$\sigma^2_- = 1 - k^2_{0\cdot 12\ldots n}$$

As we have already found that this is equal to $r^2_{0\cdot 12\ldots n}$ [formula 272] we have

$$\sigma_- = r_{0\cdot 12\ldots n} \quad \text{(Standard deviation of estimated standard scores is equal to the multiple correlation coefficient)}.[274]$$

since $r_{0\cdot 12\ldots n}$ is of necessity positive. Total and partial correlation coefficients may be positive or negative; multiple correlation coefficients can only be positive. Thus continuing we have:

$$\begin{aligned} r^2_{0\cdot 12\ldots n} &= \frac{1}{N}\Sigma z_0 \bar{z}_0 \\ &= \frac{1}{N}\Sigma z_0 (\beta_{01\cdot 23\ldots n} z_1 + \beta_{02\cdot 13\ldots n} z_2 + \ldots + \beta_{0n\cdot 12\ldots n-1} z_n) \\ &= r_{01}\beta_{01\cdot 23\ldots n} + r_{02}\beta_{02\cdot 13\ldots n} + \ldots + r_{0n}\beta_{0n\cdot 12\ldots n-1} \\ &= \frac{1}{\Delta_{00}}[r_{01}\Delta_{01} - r_{02}\Delta_{02} + r_{03}\Delta_{03} - \ldots (-1)^n r_{0n}\Delta_{0n}]. \end{aligned}$$

Referring to the major determinant, we see that, expanding it in terms of the elements of the first column, it is given by

$$\Delta = \Delta_{00} - r_{01}\Delta_{01} + r_{02}\Delta_{02} - \cdots + (-1)^n r_{0n}\Delta_{0n}$$

thus

$$r^2_{0\cdot 12\ldots n} = \frac{1}{\Delta_{00}}(\Delta_{00} - \Delta) = 1 - \frac{\Delta}{\Delta_{00}}$$

or

$$r_{0\cdot 12\ldots n} = \sqrt{1 - \frac{\Delta}{\Delta_{00}}} \qquad \text{(Determinantal solution of the multiple correlation coefficient)} \ldots\ldots [275]$$

and further

$$k_{0\cdot 12\ldots n} = \sqrt{\frac{\Delta}{\Delta_{00}}} \qquad \text{(Determinantal solution of the multiple alienation coefficient)} \ldots\ldots [276]$$

As a corollary to the two derivations [formulas 271 and 276] we have

$$\sqrt{\frac{\Delta}{\Delta_{00}}} = k_{01\cdot 23\ldots n}\, k_{02\cdot 34\ldots n} \times \ldots \times k_{0n} \qquad \ldots\ldots [277]$$

The preferable method for calculating $k_{0\cdot 12\ldots n}$ depends upon the order and whether the partial alienation and correlation coefficients are needed in the solution of the particular problem.

The theoretical solution of the n-variable problem is now complete except for the probable errors of the constants involved. The standard errors of certain constants may be immediately written down by analogy with the usual two variable situations, simply noting, e.g., that $x_{0\cdot 2}$ replaces x_0 and $x_{1\cdot 2}$ replaces x_1, etc. Thus we have by parity with formula [108 b]

$$\sigma_{r_{01\cdot 2}} = \frac{k^2_{01\cdot 2}}{\sqrt{N}} \qquad \text{(Standard error of a partial coefficient of correlation, 3 variables)} \ldots\ldots [278]$$

$$\sigma_{r_{01\cdot 2}} = \frac{k^2_{0\cdot 12}}{\sqrt{N}} \qquad \text{(Standard error of a multiple coefficient of correlation, 3 variables)} \ldots\ldots [279]$$

By parity with formula [107]

$$\sigma_{b_{01\cdot 2}} = \frac{\sigma_{0\cdot 2}\, k_{01\cdot 2}}{\sigma_{1\cdot 2}\sqrt{N}} = \frac{\sigma_{0\cdot 12}}{\sigma_{1\cdot 2}\sqrt{N}} \qquad \text{(Standard error of a } b \text{ regression coefficient, } -3 \text{ variables)} \ldots\ldots [280]$$

Plainly we may, in the case of n independent variables, deal with residuals of higher order just as we have with residuals of first and zero order and obtain:

$$\sigma_{r_{01\cdot 23\ldots n}} = \frac{k^2_{01\cdot 23\ldots n}}{\sqrt{N}} \qquad \text{(Standard error of a partial coefficient of correlation)} \ldots\ldots [281]$$

$$\sigma_{r_{0.123\ldots n}} = \frac{k^2_{0.123\ldots n}}{\sqrt{N}} \quad \text{(Standard error of a multiple coefficient of correlation)} \ldots\ldots\ldots\ldots [282]$$

$$\sigma_{b_{01.23\ldots n}} = \frac{\sigma_{0.123\ldots n}}{\sigma_{1.23\ldots n}\sqrt{N}} \quad \text{(Standard error of a regression coefficient)} \ldots\ldots\ldots\ldots\ldots [283]$$

Section 84. The Method of Successive Approximations

With more than five variables either of the preceding methods is laborious, and to meet this situation I have developed and herewith present a method of successive approximations to the values of the regression coefficients and to the multiple correlation coefficient. I have not as yet developed other than empirical tests of convergency. The method may be best presented in connection with a numerical illustration.

If given all the regression coefficients except the first, we may write

$$\bar{z}_0 = w_1 z_1 + \beta_{02.13\ldots n} z_2 + \beta_{03.124\ldots n} z_3 + \cdots + \beta_{0n.12\ldots n-1} z_n \ldots [284]$$

in which w_1 is unknown, but all the β's are known. We may now determine w_1. Designating the right-hand member, i.e., the total right-hand composite inclusive of $w_1 z_1$ by c and the right-hand composite exclusive of $w_1 z_1$ by $(c - 1)$ [to be read, "the composite exclusive of variable 1"] we have

$$\bar{z}_0 = w_1 z_1 + (c - 1) \ldots\ldots\ldots\ldots\ldots\ldots [284\,a]$$

The problem is now a simple three variable problem, the variables being z_0, z_1 and $(c - 1)$ the correlations between which we will designate as r_{01}, $r_{0(c-1)}$ and $r_{1(c-1)}$. Two of these correlations have to be determined. Both $r_{0(c-1)}$ and $r_{1(c-1)}$ are correlations between one variable and a weighted sum and are given by formula [149]. Thus we immediately have the regression coefficient of z_0 upon z_1:

$$w_1 = \frac{r_{01} - r_{0(c-1)} r_{1(c-1)}}{k^2_{1(c-1)}} \ldots\ldots\ldots\ldots\ldots [285]$$

and the regression coefficient z_0 upon $(c - 1)$ equals

$$\frac{r_{0(c-1)} - r_{01} r_{1(c-1)}}{k^2_{1(c-1)}} \ldots\ldots\ldots\ldots\ldots [286]$$

The weight w_1 as thus determined must be identical with $\beta_{01.23\ldots n}$ and the regression coefficient of $(c - 1)$ as thus determined must equal 1.0 else a better fit than the regression

equation fit has been obtained, which we know is impossible. We therefore see that if we know all of the regression coefficients except one, we can determine that one without resorting to the evaluation of two lengthy determinants.

The thought occurred to me that with reasonable weightings, guesses, or weightings somehow derived from a priori considerations, for a large number of variables no one of which was of greater importance than all the rest combined, it was to be expected that the closeness of estimate of the weighted sum of all the variables but one, which I shall call $(c - 1)$, would vary less than the weight guessed for the one. Thus if the guessed weights are w_1, w_2, $w_3 \cdots w_n$, and if c is the weighted sum $(w_1z_1 + w_2z_2 + w_3z_3 + \cdots w_nz_n)$, the calculation of the regression coefficient of z_0 upon z_1, i.e., the calculation of $\beta_{01\cdot(c-1)}$ would result in a closer approach to $\beta_{01\cdot23\ldots n}$ than, in all likelihood, was w_1. We will call this regression coefficient w_{11} and take it as a second approximation to $\beta_{01\cdot23\ldots n}$. A similar procedure using w_1, w_3, w_4, . . .w_n (not w_{11}, w_3, $w_4, \cdots w_n$) will result in a second approximation w_{22} to the correct weight for z_2, etc., for each of the other variables. We then have weights w_{11}, w_{22}, $w_{33} \cdots w_{nn}$ and may repeat the process obtaining third approximation values w_{111}, w_{222}, $w_{333}, \cdots w_{nnn}$ and still other approximations should they be needed. Just as soon as the repetition of the process results in new weights which are identical with those used in obtaining them we have the proof that the regression coefficients have been found, since as pointed out (following formula 286) this is the unique property of the regression coefficients. Therefore if repeating the process a fourth time should give $w_{1111} = w_{111}$, $w_{2222} = w_{222}$, etc., we know that $w_{1111} = \beta_{01\cdot23\ldots n}$, $w_{2222} = \beta_{02\cdot13\ldots n}$, etc., and the problem is solved. We will not expect identical agreement, but such agreement as is needed for practical purposes, say within .1 per cent, .01 per cent, or whatever other limit is self-imposed. Presumably the larger the number of variables the more rapidly convergent are the successive approximations, but I am not able to supply the theoretical proof that the convergence must take place under all circumstances. A second check upon the general approximation to regression equation weightings may be found in the size of the multiple correlation

obtained. For convergence to be present this must increase for every step.

The following example which has only six variables, and therefore constitutes a more severe test than would a problem having a larger number of variables, is given. The variables are: 0, the criterion, being a measure of general scholastic success of school children in two successive elementary school grades (population about 300); the remaining variables are the scores made by the children in the five tests comprising one of the forms of the National Intelligence tests.

(1) A test in arithmetical reasoning
(2) A test in sentence completion
(3) A test in logical selections of reasons for conduct.
(4) A test in naming synonyms and antonyms.
(5) A test in substituting digits for symbols.

The correlations between scores are

TABLE LXIII

Variables

		0	1	2	3	4
Variables	1	.4017				
	2	.6003	.2332			
	3	.2379	.1986	.1747		
	4	.6807	.2569	.4520	.2628	
	5	.3553	.1064	.2139	.0033	.2989

The symbol c will stand for the composite score according to whatever weightings are used upon the five tests; the symbols $(c - 1)$, $(c - 2)$, etc., stand for the composite scores upon all five tests, except test one, except test two, etc. The problem is to make r_{0c} a maximum. Treating one of the five variables as unique and obtaining a composite score on the other four, gives us a three variable problem, the variables being 0, u, $(c - u)$ in which u stands for the unique variable, being in turn 1, 2, 3, 4, 5, and the regression equation being

$$\bar{z}_0 = \beta_{0u.(c-u)}\, z_u + \beta_{0\,(c-u).u}\,(c - u) \quad \ldots\ldots\ldots\ldots [287]$$

The value of the second regression coefficient will ordinarily be in the neighborhood of 1.00, but it does not enter into our

present treatment. The first regression coefficient is the new weight w_{uu}, determined for z_u and is given by

$$w_{uu} = \frac{r_{ou} - r_{o\,(c-u)} r_{u\,(c-u)}}{k^2_{u\,(c-u)}} \quad \ldots\ldots\ldots\ldots [288]$$

Let s stand for the sum of the products of the correlations of the independent variables with the criterion into the weights of the independent variables, i.e.,

$$s = w_1 r_{01} + w_2 r_{02} + w_3 r_{03} + w_4 r_{04} + w_5 r_{05} \ldots\ldots\ldots [289]$$

Let S stand for twice the sum of all product terms of the sort $w_u w_{u'} r_{uu'}$, i.e., S in our present problem is a summation of 2×10 terms as follows:

$$S = 2\,(w_1 w_2 r_{12} + w_1 w_3 r_{13} + w_1 w_4 r_{14} + w_1 w_5 r_{15} + w_2 w_3 r_{23} + w_2 w_4 r_{24} + w_2 w_5 r_{25} + w_3 w_4 r_{34} + w_3 w_5 r_{35} + w_4 w_5 r_{45}) \ldots\ldots\ldots\ldots [290]$$

Let $2\,S_u$ stand for the sum of those terms in S (2×4 in number in our present problem) which involve w_u. Thus S is equal to the sum of the S_u, or in the present problem,

$$S = S_1 + S_2 + S_3 + S_4 + S_5 \ldots\ldots\ldots\ldots [290\,a]$$

and finally let Sw^2 stand for the sum of the squares of the weights. That is,

$$Sw^2 = w^2_1 + w^2_2 + w^2_3 + w^2_4 + w^2_5 \ldots\ldots\ldots\ldots [291]$$

We readily obtain by formulas [163] and [149]

$$\sigma_c = \sqrt{Sw^2 + S} \quad \text{(Standard deviation of the } c \text{ composite score)} \ldots\ldots\ldots\ldots [292]$$

$$r_{oc} = \frac{s}{\sigma_c} \quad \text{(Correlation of criterion with the } c \text{ composite score)} \ldots\ldots\ldots\ldots [293]$$

$$\sigma_{c-u} = \sqrt{Sw^2 + S - w^2 u - 2\,S_u} \quad \text{(Standard deviation of the } c\text{–}u \text{ composite score)} \ldots [294]$$

$$r_{o\,(c-u)} = \frac{s - w_u r_{ou}}{\sigma_{(c-u)}} \quad \text{(Correlation of the criterion with the } c\text{–}u \text{ composite score)} \ldots\ldots\ldots\ldots [295]$$

$$r_{u\,(c-u)} = \frac{S_u}{w_u \sigma_{c-u}} \quad \text{(Correlation of the test treated uniquely with the } c\text{–}u \text{ composite score)} \ldots\ldots\ldots [296]$$

It will be noted that if we have a problem involving one dependent variable and n independent variables that there are n terms in s, $n(n-1)$ terms in S, $(n-1)$ terms in S_u. We now have all the requisite formulas and may proceed with the calculation. For our first series of weights we will take $w_1 = 2$, $w_2 = 4$, $w_3 = 1$, $w_4 = 5$ and $w_5 = 2$, which are roughly proportional to the total correlation coefficients of the tests with the criterion. In the accompanying table p stands for the variable designated in the stub.

TABLE LXIV

Variables

			0	1	2	3	4	5	
		Wts.	$w_p r_{0p}$	$w_1 w_p r_{1p}$	$w_2 w_p r_{2p}$	$w_3 w_p r_{3p}$	$w_4 w_p r_{4p}$	$w_5 w_p r_{5p}$	(Wts.)²
	1	2	.8034		1.8652	.3972	2.5690	.4256	4.
	2	4	2.4012	1.8652		.6988	9.0400	1.7112	16.
Variables	3	1	.2379	.3972	.6988		1.3140	.0066	1.
	4	5	3.4035	2.5690	9.0400	1.3140		2.9890	25.
	5	2	.7106	.4256	1.7112	.0066	2.9890		4.
									$50 = Sw^2$
			7.5566	5.2570	13.3152	2.4166	15.9120	5.1324	42.0332 = S
			$= s$	$= S_1$	$= S_2$	$= S_3$	$= S_4$	$= S_5$	
$\sigma^2_{c-u} = \sigma^2_c - w^2_u - 2S_u$			= 77.5192	49.4028	86.2000	35.2092	77.7684	92.0332 =	
σ_{c-u}			= 8.8090	7.0286	9.2844	5.9337	8.8186	$S + Sw = \sigma^2_c$	
$r_{0(c-u)} = \dfrac{s - w_u r_{ou}}{\sigma_{c-u}}$			= .7660	.7334	.7882	.7000	.7768	$\sigma_c = 9.59339$	
$r_{u(c-u)} = \dfrac{S_u}{w_u \sigma_{c-u}}$			= .2982	.4736	.2602	.5360	.2910	$r_{0c} = s/\sigma_c$	
									$= .7877$
Second approximation									
Wts. $w_{uu} = \beta_{0u(c-u)}$			= .1905	.3265	.0350	.4285	.1412		

We will later find that no one of these weights is in error by as much as 0.1, but that is anticipating. Using these new weights we repeat the process as shown in the following table.

TABLE LXV

Variables

			0	1	2	3	4	5	
		Wts.	$w_p r_{0p}$	$w_{11} w_p r_{1p}$	$w_{22} w_p r_{2p}$	$w_{33} w_p r_{3p}$	$w_{44} w_p r_{4p}$	$w_{55} w_p r_{5p}$	(Wts.)^2
Variables	1	.19	.07632		.01462	.00113	.02099	.00283	.0361
	2	.33	.19810	.01462		.00173	.06414	.00988	.1089
	3	.03	.00714	.00113	.00173		.00339	.00001	.0009
	4	.43	.29270	.02099	.06414	.00339		.01799	.1849
	5	.14	.04974	.00283	.00988	.00001	.01799		.0196
									.3504
			.62400	.03957	.09037	.00626	.10651	.03071	.27342
$\sigma^2_{c-u} = \sigma^2_c - w^2_{uru} - 2 S_u$			= .50858	.33418	.61040	.22590	.54281		
σ_{c-u}			= .71315	.57808	.78229	.47529	.73676		
$r_{0(c-u)} = \frac{s - w_{uru} r_{ou}}{\sigma_{c-u}}$			= .767975	.736745	.789549	.697049	.779444	σ^2_c = .62382	
								σ_c = .789823	
$r_{u\,(c-u)} = \frac{S_u}{w_u \sigma_{c-u}}$			= .29203	.47372	.26674	.52115	.29773	$r_{0c} = s/\sigma$	
								= .790051	
Third approximation									
Wts. $w_{uruu} = \beta_{0u.(c-u)}$			= .1940	.3234	.0294	.4358	.1352		

Tabulating the results thus far obtained, we have

TABLE LXVI

Variables	Weights First Guess	Weights Second Approximation	Weights Third Approximation
1	2	.19	.1940
2	4	.33	.3240
3	1	.03	.0294
4	5	.43	.4358
5	2	.14	.1352
Multiple correlation resulting	.7877	.79005	

The first weights give a multiple correlation of .7877 and lead to the determination of the second approximation weights. The second weights give a multiple correlation of .79005 and lead to the determination of the third approximation weights. The third weights differ so slightly from the second that for ordinary purposes one would stop the calculation here, use the third weights as final and take the multiple correlation as equal to .7901 since it will be a trifle above .79005. The method of calculation of the weights here shown involves but a fraction of the time necessary to evaluate the determinants necessary to a solution. This is true for three reasons:

(*a*) Number of operations is much smaller.

(*b*) No checking for inaccuracies in any of the calculations, except that for the last weights derived, need be made, as a small error leading to a wrong approximate weight will be corrected in the next step.

(*c*) Partial regression coefficients $\beta_{0u.(c-u)}$, except for the last step where greater accuracy may be desired, may be made by the aid of the alignment chart.

A further device which is serviceable is to compare r_{0c} with each of the $r_{0(c-u)}$ values in the same calculation. Should any one of the $r_{0(c-u)}$ correlations be larger than r_{0c} it indicates that the weight used for the test in question is worse than would be a weight of zero. Referring to the first of the calculations above, we find that $r_{0c} = .7877$ and that $r_{0(c-3)} = .7882$. This means that the weight which was assumed for test 3, namely 1.00, is a worse weight than would be the weight zero. Thus if the problem is such that only positive weights have been used as the first approximations, any variables

which should have negative weights will probably be discovered in the first calculation by the correlation $r_{0(c-u)}$ turning out higher than r_{0c}.

The solution by determinants of the above problem correct to seven decimal places has been kindly supplied to me by Miss Ella Woodyard.

$w_1 = .19412341$ $\quad w_4 = .43693997$
$w_2 = .32392693$ $\quad w_5 = .13466545$
$w_3 = .02748474$ $\quad r_{0\cdot 12345} = .79009053$

It will be seen that the maximum error in the third approximation weights is .0019, which is the error for w_3. This would probably be considered a negligible error. Should, however, greater accuracy be required, a determination of fourth order approximation weights will give it. Actually such calculation gives weights, no one of which is in error by more than .0001. I have also made a fifth calculation resulting in the multiple correlation $r_{0\cdot 12345} = .79009038$ which is seen to be in error by .00000015. Thus for these data there can be no doubt that rapid convergence actually exists. One desiring to practice the method is referred to Yerkes (1921), where abundant multiple correlation equation material already worked out by the determinantal method is to be found. I have used this method upon a variety of problems and have always found convergence. Much time will be saved if the original guess as to the final weights are excellent, but the method does not require approximate accuracy in the original weights. To illustrate this, let us work the present problem, starting with weights 0, 1, 2, − 2, − 1, which are about as unreasonable as it is possible to assume. The calculation gives

TABLE LXVII

Variables	Weights First Guess	Weights Second Approximation	Weights Third Approximation
1	0	.4	.188
2	1	.5	.332
3	2	.2	.031
4	− 2	.7	.437
5	− 1	.3	.151
Multiple correlation resulting	− .23	.784	

Evidence of convergence is not clearly apparent from these three series of weights, but it of course is apparent by comparison of the third weights with the correct values. The very poor choice of original weights has increased the number of calculations necessary to establish convergence, but it has had no other effect.

A possible difficulty in the calculation of the $\beta_{0u\cdot(c-u)}$ coefficients in case one of the approximate weights is zero may be mentioned. In case $w_u = 0$,

$$r_{u(c-u)} = \frac{S_u}{w_u \sigma_{c-u}} = \frac{0}{0} \quad \ldots\ldots\ldots\ldots\ldots\ldots [296\,a]$$

To avoid this indeterminate form we may write

$$r_{u(c-u)} = \frac{S w_p r_{up}}{\sigma_{c-u}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots [297]$$

instead of the preceding, which is generally shorter to use. As an illustration of this situation it may be noted that w_1 was chosen equal to 0 in Table LXVII. Thus $S_1 = 0$ and $r_{1(c-1)} = \frac{0}{0}$ by formula [296]. Using formula [297] we have

$$r_{1(c-1)} = \frac{w_2 r_{12} + w_3 r_{13} + w_4 r_{14} + w_5 r_{15}}{\sigma_{c-1}} = \frac{.0102}{2.7465} = .0037$$

This is no longer indeterminate. Except in this calculation of $r_{u(c-u)}$ no special procedure will be necessary on account of a zero weight. The introduction of zero weights where reasonable leads to a simplification of the numerical work. For the problem in hand, if the first estimated weights had been 2, 4, 0, 5, 2 instead of 2, 4, 1, 5, 2 it would have simplified the first calculation and led to rapid convergence. It is well to estimate a zero weight whenever in doubt. The regression weights as just determined are of course β coefficients, $w_1 = \beta_{01\cdot 23\ldots n}$, $w_2 = \beta_{02\cdot 13\ldots n}$, etc., pertaining to the equation [260]

$$\bar{z}_0 = \beta_{01\cdot 23\ldots n} z_1 + \beta_{02\cdot 13\ldots n} z_2 + \cdots + \beta_{0n\cdot 12\ldots n-1} z_n$$

Making the substitutions of equations [247] and [248] immediately gives the regression equation involving gross scores

$$\overline{X}_0 = b_{01\cdot 23\ldots n} X_1 + b_{02\cdot 13\ldots n} X_2 + \cdots + b_{0n\cdot 12\ldots n-1} X_n + c$$

The regression coefficients and the multiple correlation coefficient are given by this successive approximation method. The partial alienation and correlation coefficients, as well as the important standard errors, may all be obtained by formulas given earlier in this chapter.

CHAPTER XII

STATISTICAL TREATMENT OF SUNDRY SPECIAL PROBLEMS

Section 85. Statistical Constants Determined from Mutilated Distributions

If a portion only of a distribution is available it is possible to reconstruct the entire distribution when a reasonable assumption of the form of the entire distribution can be made. The principle is applicable to any form, but only in case the assumed form is normal are the constants enabling a ready calculation available in tables. Let us assume that data for the tail of a sharply truncated distribution, which is in truth normal, are available. The "tail" may be greater or less than one-half of the total or untruncated distribution. The distance from the stump to the mean of the tail bears a ratio to the standard deviation of the tail which changes as the point of truncation changes; conversely, the value of this ratio determines the proportion of the total distribution which is represented by the tail. This is the property utilized by Pearson and Lee (1908), and by Lee (1914), in reconstructing the total distribution from a sharply truncated portion. Tables facilitating this process are to be found in the references cited.

There are other properties, such as the ratio between the median deviation and the mean deviation of the tail measured from the point of truncation, which can be utilized to the same purpose, and it is not at all evident that the error of such determination is greater than that of the Pearson and Lee determination. The probable errors which establish the reliability of either method are at present unavailable. The accompanying Table, LXVIII, gives the ratio of the median deviation from the stump, to the mean deviation, for successive percentages of a total normal distribution.

TABLE LXVIII

q	$\frac{\text{Median}}{\text{Mean}}$	q	$\frac{\text{Median}}{\text{Mean}}$	q	$\frac{\text{Median}}{\text{Mean}}$
.01	.7363	34	.8143	67	.8833
2	.7425	35	.8162	68	.8858
3	.7470	36	.8181	69	.8884
4	.7508	37	.8199	70	.8909
5	.7541	38	.8218	71	.8935
6	.7571	39	.8237	72	.8962
7	.7599	40	.8256	73	.8988
8	.7625	41	.8276	74	.9016
9	.7650	42	.8295	75	.9043
10	.7674	43	.8314	76	.9071
11	.7697	44	.8334	77	.9100
12	.7719	45	.8353	78	.9129
13	.7741	46	.8373	79	.9159
14	.7762	47	.8393	80	.9189
15	.7782	48	.8413	81	.9220
16	.7803	49	.8433	82	.9252
17	.7823	50	.8453	83	.9284
18	.7843	51	.8474	84	.9317
19	.7862	52	.8495	85	.9350
20	.7881	53	.8516	86	.9384
21	.7901	54	.8537	87	.9420
22	.7920	55	.8558	88	.9456
23	.7938	56	.8580	89	.9492
24	.7957	57	.8601	90	.9530
25	.7976	58	.8623	91	.9569
26	.7995	59	.8646	92	.9610
27	.8013	60	.8668	93	.9651
28	.8032	61	.8691	94	.9694
29	.8050	62	.8714	95	.9738
30	.8069	63	.8737	96	.9785
31	.8087	64	.8761	97	.9833
32	.8106	65	.8785	98	.9884
33	.8125	66	.8809	99	.9939

Entering Table LXVIII with the ratio given by the data leads to q, the proportion in the tail, and thus to N, the population of the total untruncated distribution. The further steps in the solution will be obvious from the problem discussed in the next paragraph.

It not unfrequently happens that the total population is known, so that the items available are (*a*) q, the proportion in the tail, (*b*) the point of truncation, and (*c*) the distribution of the tail measures. In this case the fitting of an assumed normal distribution is very simple. Let m = the mean of the tail measured from the stump; let D = the distance from the mean of the total distribution to the stump; let σ = the standard deviation of the total distribution; and let x and z have the values of Table K–W when entered with the argument q. We then have, from formula [53]

$$x = \frac{D}{\sigma}, \text{ or } D = x\sigma \quad \ldots\ldots\ldots\ldots\ldots\ldots [298]$$

$$\frac{z}{q} = \frac{D}{\sigma} + \frac{M}{\sigma} \text{ or } \sigma = \frac{M}{\frac{z}{q} - x} \quad \ldots\ldots\ldots\ldots\ldots [299]$$

Solving these two equations for σ and D completes the problem.

As an illustration of the use of Table LXVIII we may calculate, from the data of Table LXIX, the constants of the total grade distribution of 15-year olds knowing the grade distribution of the portion found in the elementary school. The children represented range from 13.5 to 14.5 years of age. We will assume that the total grade distribution is normal and that the elementary school portion is a sharply truncated tail, though in case the compulsory school attendance law applies only to the elementary school this assumption is undoubtedly in error, leading to a larger estimate of the number in the high school than would actually be found there. In the grade scale used, 3.0 means the beginning of the third grade, 3.25 the middle of the low third, 3.75 the middle of the high third, etc.

TABLE LXIX

Grade Distribution of 14-Year Olds Obtained from Certain Virginia Survey Data

Grade	3.25	3.75	4.25	4.75	5.25	5.75	6.25	6.75	7.25	7.75	8.25	8.75	Total
Number of Pupils	1	2	4	7	13	11	61	60	82	96	40	34	411

The point of truncation is 9.00. Calculation gives

Mdn measured from 9.00 = − 1.685

M measured from 9.00 = − 1.835

$$\frac{\text{Mdn}}{M} = .9181$$

From Table LXVIII, $q = .7975$. This proportion is represented by 411 pupils, so that the number in the untruncated or total distribution is 515 pupils. The standard deviation of the total distribution is, by formula [299], equal to 1.521 grades, and D, the distance from the stump to the mean of the total distribution, is found by formula [298] to equal 1.266 grades. Accordingly the constants of the untruncated distribution of fourteen-year olds are

Mean grade = 7.734
Standard deviation = 1.521 grades
Population = 515 pupils

Section 86. Correlation Determined from Mutilated Distributions

The ability to determine the constants of a total distribution from a known fraction of it may be turned to practical account in decreasing the size of populations necessary for an assigned accuracy. The procedure may be illustrated by a problem, the data for which have been kindly supplied by Miss Margaret V. Cobb.

TABLE LXX

Numbers of Pupils Obtaining Designated Scores upon a Symbol-Digit Substitution Test

Test Scores	School Grades			
	4.25	4.75	8.25	8.75
105			4	6
100			4	5
95		1	1	3
90		1	7	3
85			5	1
80	1			3
75	3	1	2	
70	1		1	
65		1		
60	3	1		
55	4	3	1	
50	3	2		
45	4	2		
40	3	1		
35	2			
30	2			
25	1	1		
20	1			
	28	14	25	21 $N = 88$

The problem which we will set is, in outline, to (*a*) calculate r from this mutilated table, (*b*) determine R, the correlation to be expected in a range of two grades, let us say the fifth and sixth, (*c*) determine the probable error of R as thus found, (*d*) determine the probable error of an R of the same size (designated R') if found from a population of the same size in grades 5 and 6, and (*e*) by comparing the reliability of R and R' endeavor to ascertain whether an artificial selection of original data will decrease the populations necessary to secure a desired reliability.

Letting school grade be the first variable, and test score the second, we find $r_{12} = .827$.

If we can determine Σ_1/σ_1, where Σ_1 is the standard deviation of the 5 and 6 grade distribution, and σ_1 that of the 4 and 8 grade distribution, we may use formula [86] to obtain R. Assuming that there are the same number, f, of pupils in each grade we have the two following distributions:

4 and 8 grade distribution	Grades	4.25	4.75	8.25	8.75	giving $\sigma^2_1 = 4.0625$ grades
	Frequencies	f	f	f	f	
5 and 6 grade distribution	Grades	5.25	5.75	6.25	6.75	giving $\Sigma^2_1 = .3125$ grades
	Frequencies	f	f	f	f	

from which the ratio $\Sigma_1/\sigma_1 = .27735$. Having this ratio and r_{12} we find by formula [186] that $R = .378$. Thus the correlation in a two grade range is rather low.

By formula [108 *b*], $\sigma_r = .317/\sqrt{N}$, but this is too small a value, as the distributions with which we are working are far from mesokurtic. Estimating the β_2's for the school grade and the test score distributions to be 1.06 and 1.94 respectively gives by formula [108 *a*], $\sigma_r = .515\sqrt{N}$, which is the preferable value in the case of this platykurtic correlation surface.

If the assumption of form of grade distribution can be made with great certainty, so that we may consider no error to enter into the ratio Σ_1/σ_1 we may obtain the standard error of R knowing that r. Starting with formula [187] and taking logarithmic differentials we have,

$$\frac{dr}{r} - \frac{dk}{k} = \frac{dR}{R} - \frac{dK}{K}$$

$$dk = d\sqrt{1 - r^2} = \frac{-r\,dr}{k}, \text{ and } dK = \frac{-RdR}{K}$$

Substituting these values for dk and dK, squaring, summing, dividing by the population, and extracting the square root, gives

$$\frac{\sigma_r}{rk^2} = \frac{\sigma_R}{RK^2}$$

or

$$\sigma_R = \sigma_r \frac{RK^2}{rk^2}$$ (Standard error of the correlation coefficient inferred from a coefficient obtained in a different range) .....[300]

Using this formula we find for the data in hand,

$$\sigma_R = 1.2387\,\sigma_r = .638/\sqrt{N} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (a)$$

Had the correlation been directly determined from the 5 and 6 grade distribution, its value would presumably be about the same $R_1 = .378$, but its standard error would have been different. Estimating the β_2's to be 2.1 and 3.0, instead of 1.06 and 1.94, as above, the standard error by formula [108 *a*] is

$$\sigma_{R'} = .873/\sqrt{N} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (b)$$

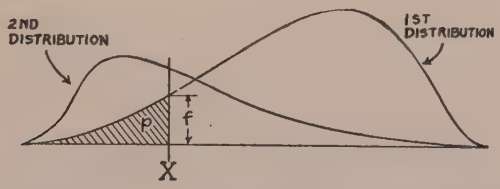
Choosing such an N for formula (*b*) as to result in the same standard error as given by formula (*a*) shows that 1.87 N are needed in the narrow 5 and 6 grade calculation to obtain an equally reliable result to that deduced for these grades by the 4 and 8 grade calculation based on N.

One cannot generalize and say that, given equal populations, more reliable results are always obtained from the wider range determination, but this is true if correlations are low, in the narrow range and not very high in the wide range — say under .40 in the former and not over .70 in the latter. If entire freedom in choosing the range of talent to be examined is present, excellent results may be expected if a fairly mesokurtic distribution, yielding a correlation between .60 and .70, can be selected, and then estimating the correlation for greater and lesser ranges by formula [186].

Section 87. The Probable Error of Percentage Measures of Overlapping

The probable error of the proportion in one distribution which exceeds or falls short of a certain percentile in a second distribution is a function of both distributions. Let the constants of the first distribution (to the right in the accompany-

ing figure) be designated by lower case letters and those of the second distribution by capitals. Let p = the proportion of the first distribution falling short of the percentile X, of the



second distribution. A change in p may be produced either by a change in X, or by a change in the proportion in the first distribution below an assigned point.

Let δ = a small change in the proportion p due to fluctuation in the second distribution.

Let d = a small change in the proportion p due to fluctuation in the first distribution.

Let Δ = a small change in the proportion p due to fluctuations in both distributions.

Then $\Delta = \delta + d$, and σ_Δ, identical with σ_p, is the standard error desired.

$$\Sigma\Delta^2 = \Sigma\delta^2 + \Sigma d^2 + 2\,\Sigma\delta d$$

Since δ and d are functions of two independent distributions they are uncorrelated and $\Sigma\delta d = 0$, so that

$$\sigma^2_p = \sigma^2_\delta + \sigma^2_d \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots [301]$$

σ_d is the standard deviation of the proportion of measures in the first distribution below the point X and by formula [40] $\sigma_d = \sqrt{pq/n}$.

If the ordinate of the first distribution per unit base at the point X is f and if the distribution is assumed sufficiently flat at this point that a small change to the right in X would pass over approximately the same number of cases as an equal change to the left, then a small change D in X causes a change of fD in the number of cases, np, of the first distribution lying below the point X. Dealing with proportions, p is affected to the extent fD/n. In consequence,

$$\sigma_\delta = \sigma_{\frac{fD}{n}}$$

In this equation f and n are constants, for we are considering fluctuation due to variability in the second distribution, so that

$$\sigma_\delta = \frac{f}{n}\sigma_D \quad \text{(See problem 7, Chapter 4)}$$

σ_D is simply the standard error of a percentile. We have by formula [42]; letting P = the proportion of the second distribution determining the point X; $Q = 1 - P$; i_P = the number of units in the class interval in which X lies; f_P the frequency of this class; and N the population of the second distribution;

$$\sigma^2_D = \frac{i^2_P\, NPQ}{f^2_P}$$

Making the proper substitutions in [301] results in

$$\sigma^2_p = \left(\frac{f^2}{n^2}\right)\left[\frac{N\, i^2_P\, PQ}{f^2_P}\right] + \left(\frac{pq}{n}\right)$$

(Square of the standard error of the proportion of a distribution falling short of or exceeding an assigned percentile of a second distribution) .....[302]

Note that in this formula the constants in ( ) refer to the first distribution and those in the [ ] to the second distribution.

If the proportion exceeding the median of the second distribution is being determined, $P = Q = \frac{1}{2}$; and if, further, the second distribution is normal, $f_P/i_P = .3989N/\Sigma$, in which Σ is the standard deviation of the second distribution, so that

$$\sigma^2_p = 1.57080\,\Sigma^2\,\frac{f^2}{Nn^2} + \frac{pq}{n}$$

(Square of the standard error of the proportion of a distribution falling short of or exceeding the median of a second and normal distribution) .....[303]

In case both distributions are normal and have the same populations and standard deviations, Table LXXI when multiplied by $1/\sqrt{N}$ gives the standard errors, in the second column, and the probable errors, in the third column, for different values of p.

In illustration of the use of Table LXXI the following problem is given: In a certain fifth grade only 40 per cent of the pupils exceed in a reading test 50 per cent of the fourth grade. We will assume the same number of pupils, 36, in each grade. What are the chances that the true test ability of the fifth grade is above that of the fourth grade? Referring to Table LXXI we find that the standard error of the proportion, .40, is

$(.689/\sqrt{36} =)$ .115. Thus the difference between the obtained proportion and the proportion in case of equally able classes, namely .10 is $(.10/.115 =)$ .87 standard errors. Entering Table K–W with $x = .87$ we obtain $q = .19$, or, in other words, the chances are 19 in 100 that the fifth grade ability is in truth as great as that of the fourth grade.

TABLE LXXI

Proportion lying Below or Above Median of Second Distribution	$\sqrt{N} \times$ the Standard Errors of the Proportions of One Distribution Below, or Above, the Median of a Second	$\sqrt{N} \times$ the P. E.'s
.001	.032	.022
.01	.105	.071
.02	.153	.103
.05	.252	.170
.10	.372	.251
.15	.462	.311
.20	.532	.359
.25	.588	.396
.30	.622	426
.35	.666	.449
.40	.689	.465
.45	.703	.474
.50	.707	.477

If, for this same problem, fourth and fifth grade means are calculated and the probable error of the difference between means found by formula [140] we will finally obtain the result that there are 14.4 chances in 100 that the fifth grade ability is in truth as great as that of the fourth grade. Thus slightly more definite results may be obtained by finding the differences between means instead of the percentage of overlapping. Formula [166] of Section 59 provides the correction for the error in a measure of overlapping due not as here to size of population but to inaccuracy in the instrument of measurement.

Section 88. A Criterion for the Addition or Elimination of Elements Having Fixt Weightings

In many trade, education, and intelligence tests, and in combining stock quotations to determine general trends, it is frequently required, because of the necessity for maintaining simplicity of procedure, to include an item in a composite at a given weight, or to reject it in toto, i.e., no adjusting of the

weight to the importance of the item is possible. A criterion for the inclusion or rejection of an item is needed for the handling of this problem.

To make the problem specific let us suppose that a questions, each scored right or wrong, are being evaluated with reference to their excellence as a ten-year-old general intelligence test battery (such, for example, are the Binet type of questions). The correlations of each of the a questions with an independent general intelligence measure and the intercorrelations between the questions constitute the requisite basic data. Having these and using the weights that are imposed, calculate correlations exactly as in the row labeled "$r_{0(c-u)}$" in Table LXIV. The highest of these correlations locates the question which contributes least. This question may be discarded and the process repeated with the $(a - 1)$ remaining questions, etc., until the number desired for the final battery are left. At each step in this process a comparison of the $r_{0(c-u)}$ correlation with the r_{0c} correlation shows how much loss, if any, in multiple correlation results from discarding the question, thus making available all the information pertinent to the problem.

All correlations should be by the usual product-moment method, even though but two degrees of merit are possible. For the intercorrelations formula [214] may be used.

Section 89. Trade Test Calibration

A procedure of evaluation, or, "calibration," of trade test questions, based upon the slope of an ogive curve, has been practiced by the Army Trade Test staff. As an illustration let us suppose questions A, B, C, and D have been correctly answered by varying proportions of unskilled and skilled artisans as shown in the following Table:

TABLE LXXII

Percentages Answering Correctly

	Novices	Apprentices	Journeymen	Experts
Question A	10	14	18	24
Question B	2	2	51	60
Question C	20	62	70	75
Question D	2	1	14	54

I have elsewhere (Kelley, 1916, simp.), pointed out, in the case of an ogive curve in which the abscissa is a scale of difficulty, and the ordinate per cent correct responses, that uncorrelated questions of the difficulty corresponding to the point of steepest slope result in more accurate determinations of ability than a similar number of questions of a different difficulty. The principle is clearly general, and can be used to scale a question given subjects of known differences in ability just as, in reverse, it can be used to determine proficiency when given scaled questions. Thus, if ogive curves, the abscissa being Novice-apprentice-journey-expert and the ordinate per cent correct responses, be plotted for each question, the steepest part of the curve will lie between the two groups most decisively differentiated from each other by the question.

Inspection shows that question A is not satisfactory either as an apprentice, journeyman, or expert question; that question B is an excellent journeyman question; C an excellent apprentice question; and D a good expert question. So far as determining the trade group with reference to which a single question will be of most value the method is excellent, but it falls short, as will every method not involving intercorrelations, of what is to be desired in a method used to select a battery of questions. A combination of this procedure with that of the previous section should give good results.

Section 90. The Determination of the Cross-over Value of a Chromosome Section *

In the following treatment certain terms will be used with meanings which may be made clear by an example: If a fly showing two mutant characters, black and vestigial, is crossed to a fly showing neither of these characters, then in the back cross progeny the characters will reappear in the *original combinations*, namely black vestigial or not-black not-vestigial, in the majority of cases, but small classes of progeny will occur that are *recombinations* of the original characters, namely, they are black not-vestigial, or not-black vestigial flies.

* I am indebted to Dr. Calvin B. Bridges for the biological statement of this problem.

To explain the occurrence of these recombinations it is assumed that *crossing-over* occurs in the section of the chromosome between the loci at which the genes for these characters are situated. The gene responsible for the development of the character black is situated in a rod-like body called a *chromosome* at a definite point which is the black *locus*. Likewise the gene for vestigial is situated in that same chromosome, the "second" at a locus some distance to the right of that of black. The second chromosome is represented twice in every cell — by the chromosome from the mother carrying the genes for black and vestigial, and by the chromosome from the father carrying in these loci the gene for not-black and the gene for not-vestigial. In the production of eggs these two chromosomes, A and A', come to lie side by side and homologous sections are interchanged by *crossing-over*. Both chromosomes break in two at a corresponding point and the left part of A joins to the right part of A' and vice-versa. The cross-over occurs at random along the chromosome. Whenever one occurs between the loci of black and vestigial, a black not-vestigial and a not-black vestigial chromosome are produced and these give rise to the character *recombinations*. However, two occurrences of crossing-over may take place coincidentally between these loci and not be detected as a recombination of the characters. Again, if three cross-overs take place between these loci only simple recombination is observed. Accordingly, unless the section is so short as to preclude double crossing-over, the number of recombinations is always less than the number of cross-overs.

The first problem of the student of this subject is to determine the number of cross-overs from the number of recombinations. This problem offers certain difficulties, but for our present problem we will assume it solved by an equation of the type

$$100\,n = 100\,(R + 2\,d) \quad \text{(The cross-over value of a chromosome section)} \ldots [304]$$

in which R is the proportion of recombinations observed to take place, d the proportion of double (plus occasional triple), cross-overs, expected from previous determinations in this general chromosome region, when the proportion of recombina-

tions is R, and 100 n is the cross-over value of the section studied as given by the experiment.

The second problem is the determination of the reliability of the cross-over value determination. This offers genuine statistical difficulties due to the variability in the ratio d/R for different lengths of chromosome and for different general regions in the chromosome. I offer the following as an empirical formula, which I believe will not be far from the mark, at least as long as uncertainty as to the ratio d/R persists:

$$100\,\sigma_n = 100\,(\sigma_R + 2\,\sigma_d)$$

(Empirical formula for the standard error of the cross-over value). [305]

in which σ_R is defined by the equation

$$\sigma_R = \sqrt{\frac{R\,(1 - R)}{N}}$$

and σ_d by the equation

$$\sigma_d = \sqrt{\frac{d\,(1 - d)}{N}}$$

N in each case being the total number of flies in the experiment. Having, either by means of formula [305] or otherwise, an estimate of the standard error of a single cross-over value determination, we come to the third problem, which is:

The utilization of several direct and indirect independent determinations of the length of the same chromosome section to arrive at the most probable value.

Let 100 n_{12} = an experimentally determined cross-over value between loci 1 and 2, and let 100 σ_{12} = its standard error: and similarly for n's and σ's with other subscripts.

If a number of loci in order are X_1, X_2, X_3, X_4 and if different experiments have been conducted so that there are separate determinations of (*a*) n_{13}, (*b*) n_{12}, and (*c*) n_{23}, the problem then is to use these three determinations to arrive at the most reliable value for the distance between X_1 and X_3. We will call this most reliable value $\bar{n}_{13}$. We have two determinations of the same distance, namely, n_{13} and $(n_{12} + n_{23})$. The standard error of n_{13} is σ_{13}, and since n_{12} and n_{23} are independent determinations they are uncorrelated and the standard error of $(n_{12} + n_{23})$ is $\sqrt{\sigma^2_{12} + \sigma^2_{23}}$. To average these two distances so as to secure a distance with the minimal standard error we

must weight each inversely as the square of its standard error as proven in the next section, formula [307]. Accordingly,

$$\bar{n}_{13} = \frac{\frac{n_{13}}{\sigma^2_{13}} + \frac{n_{12} + n_{23}}{\sigma^2_{12} + \sigma^2_{23}}}{\frac{1}{\sigma^2_{13}} + \frac{1}{\sigma^2_{12} + \sigma^2_{23}}}$$

Should there be a third independent means of determination, e.g., were independent values of n_{14} and n_{34} available, the procedure would be similar, giving

$$\bar{n}_{13} = \frac{\frac{n_{13}}{\sigma^2_{13}} + \frac{n_{12} + n_{23}}{\sigma^2_{12} + \sigma^2_{23}} + \frac{n_{14} - n_{34}}{\sigma^2_{14} + \sigma^2_{34}}}{\frac{1}{\sigma^2_{13}} + \frac{1}{\sigma^2_{12} + \sigma^2_{23}} + \frac{1}{\sigma^2_{14} + \sigma^2_{34}}}$$

(The best value for a distance scaled in three independent ways: (*a*) in toto, (*b*) in parts, (*c*) in parts).....[306]

Any further number of independent determinations may be utilized in the same manner. It may happen that the number of possible means of determination is so great as to make the labor of utilizing all of them excessive, in which case certain clearly defined loci preferably between 30 and 40 units apart may be carefully determined using all the data and other points located with reference to them using data between two loci already scaled.

Section 91. The Best Weighted Average of Independent Variables

To complete the proof of the preceding section it remains to establish the theorem that the best weighted average of n independent measures of the same magnitude is that obtained by weighting each inversely as the square of its standard error.

We will first prove it for two variables, a_1 and a_2, having standard errors σ_1 and σ_2. It is required to so distribute the total weight of 1.00 between a_1 and a_2 that the standard error, σ_-, of the weighted average, $\bar{a}$, shall be a minimum. Let the weights be w_1 and w_2. We have

$$w_1 + w_2 = 1$$
$$\bar{a} = w_1 a_1 + (1 - w_1)\, a_2$$
$$\sigma^2_- = w^2_1\, \sigma^2_1 + (1 - w_1)^2\, \sigma^2_2 + 2\, w_1\, (1 - w)\, \sigma_1\, \sigma_2\, r_{12}$$

in which r_{12} is equal to zero as a_1 and a_2 are by hypothesis independent measures, so that

$$\sigma^2_- = w^2_1\, \sigma^2_1 + \sigma^2_2 - 2\, w_1\, \sigma^2_2 + w^2_1\, \sigma^2_2 ,$$

Differentiating with respect to w_1, setting the derivative equal to zero, and solving for w_1 and w_2 results in

$$\frac{w_1}{w_2} = \frac{\frac{1}{\sigma^2_1}}{\frac{1}{\sigma^2_2}}$$

$$\bar{a} = \frac{\frac{1}{\sigma^2_1} a_1 + \frac{1}{\sigma^2_2} a_2}{\frac{1}{\sigma^2_1} + \frac{1}{\sigma^2_2}}$$ (Best weighted average of two independent measures) . . . . [307]

If a third variable which is independent of the first two is included it cannot change the best relative weightings of the first two, therefore

$$\frac{w_1}{w_2} = \frac{\frac{1}{\sigma^2_1}}{\frac{1}{\sigma^2_2}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (a)$$

must still hold, and by parity

$$\frac{w_1}{w_3} = \frac{\frac{1}{\sigma^2_1}}{\frac{1}{\sigma^2_3}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (b)$$

$$\frac{w_2}{w_3} = \frac{\frac{1}{\sigma^2_2}}{\frac{1}{\sigma^2_3}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (c)$$

Further, the sum of the weights must equal 1, that is,

$$w_1 + w_2 + w_3 = 1 \ldots\ldots\ldots\ldots\ldots\ldots (d)$$

By inspection it is seen that the weights in the following equation meet these four conditions:

$$\bar{a} = \frac{\frac{1}{\sigma^2_1} a_1 + \frac{1}{\sigma^2_2} a_2 + \frac{1}{\sigma^2_3} a_3}{\frac{1}{\sigma^2_1} + \frac{1}{\sigma^2_2} + \frac{1}{\sigma^2_3}}$$ (Best weighted average of three independent measures) . . . . . [308]

Having four conditions to meet and but three weights this solution is unique. It is obvious from the steps involved that the proof may be extended to cover any number of variables, so that in general

$$\bar{a} = \frac{\frac{1}{\sigma^2_1} a_1 + \frac{1}{\sigma^2_2} a_2 + \frac{1}{\sigma^2_3} a_3 + \cdots + \frac{1}{\sigma^2_n} a_n}{\frac{1}{\sigma^2_1} + \frac{1}{\sigma^2_2} + \frac{1}{\sigma^2_3} \cdots + \frac{1}{\sigma^2_n}}$$ (Best average of n independent variables) . . . . . [309]

Section 92. Psychophysical Methods

The excellent treatment of the statistical processes involved in the handling of the various psychophysical methods given in Brown and Thomson (1921) makes an exhaustive treatment here unnecessary; however, the very important process of fitting smooth curves to data collected by the "constant method," or the "method of right and wrong cases," is treated of in connection with Fechner's fundamental table of the normal probability integral. Table K–W is so much more serviceable in this connection, both because of the type of entry which it contains and because of the greater accuracy which it permits, that the process is herewith described in full.

When successive stimuli, s_1, s_2, $s_3, \cdots s_m$, are each compared a number of times, N_1, N_2, $N_3, \cdots N_m$, with a constant stimulus, k, and the subject is required to act in each case by calling the stimulus greater than or less than the constant stimulus, there results a progression of proportions, p_1, p_2, $p_3, \cdots p_m$, giving the proportion of times that each stimulus is considered greater than the standard of comparison, k. If the smallest of the variable stimuli is much smaller than k and the largest is much larger, the proportions will run from .00 to 1.00 and if plotted will give an ogive curve. If the smallest and largest stimuli are not sufficiently different from K to lead to proportions of .00 and 1.00 at the extremes, some reasonable assumption as to the distribution of these tail measures must be made. From the general nature of the ogive curves found in psychological data obtained as described, it has been surmised that the integral of a normal curve may ordinarily be taken as well representing the distribution of proportions in the tails as well as in the more central portion of the curve.

The problem is, therefore, to fit a curve of the type

$$p = \frac{1}{N}\int_{-\infty}^{\frac{s-\bar{s}}{\sigma}} y\, ds$$

to the observed data. The magnitude $\bar{s}$ is that stimulus at which p equals $\frac{1}{2}$. It is not an observed value of s, but is to be determined from all the data; s is any one of the variable stimuli; σ is the standard deviation, in terms of the units of s,

of the normal distribution of which p is the integral. Accordingly $(s - \bar{s})/\sigma$ is a deviation from the mean expressed in terms of the standard deviation, that is, it is comparable to x of Table K–W, and $y\sigma/N$ is comparable to z of that table. Assume that the ogive curve is the integral of the normal distribution

$$y = \frac{N}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{s-\bar{s}}{\sigma}\right)^2}$$

Each proportion, $p_1, p_2, p_3, \cdots p_m$, is a fraction of the area under this curve and for each such proportion there is a value $x_1, x_2, x_3, \cdots x_m$, which may be obtained from Table K–W.

Even if the values of s and σ have been determined in the best possible manner there will still be discrepancies between x_1 and $(s_1 - \bar{s})/\sigma$; x_2 and $(s_2 - \bar{s})/\sigma$; etc. due to the best fit ogive not being a perfect fit. The problem may now be restated in more specific terms. It is required to determine $\bar{s}$ and σ (in the parlance of psychology, $\bar{s}$ is the threshold and σ is the dispersion of the measures giving the threshold), so that the sum of the squares of the deviations, $[x - (s - \bar{s})/\sigma]$, shall be a minimum.

In the early statement of the problem by Fechner and Müller it was argued that the sum of the squares of the deviations of the obtained p's from calculated p's should be made a minimum, but as Urban (1909), (1912) and Thomson (1919 dir.), have shown this is plainly in error and the deviations in the x's, as indicated above, are the proper ones to treat by the method of least squares.

For each proportion p there is an x which differs from $(s - \bar{s})/\sigma$ by a certain amount, and the standard error of this difference is identical with the standard error of the x, for s has no error in it, being a given stimulus. If, therefore, the standard errors of $x_1, x_2, x_3, \cdots x_m$ are obtained, we know exactly what weights to give to the m derivations in arriving at the best values of $\bar{s}$ and σ, for by the theorem of the preceding section, independent measures of unequal reliability should be weighted inversely as the squares of their standard errors.

The deviation x is simply a percentile value, and the standard error of a percentile [43] has been shown to be equal to

$$\frac{\sigma}{N}\sqrt{\frac{pq}{z^2}}$$

Accordingly the m residuals, $[x_1 - (s_1 - \bar{s})/\sigma]$, $[x_2 - (s_2 - \bar{s})/\sigma]$, $\cdots$ must be weighted

$$\frac{N_1 z^2_1}{\sigma^2 p_2 q_1}, \frac{N_2 z^2_2}{\sigma^2 p_2 q_2}, \cdots$$

respectively. Since σ^2 is a constant for the entire procedure, it may be dropped without affecting the relative weightings. N_1, N_2, $\cdots$ depend upon the particular experiment. The remainder is z^2/pq and is the product of the entries in the z/p and z/q columns of Table K–W. Except for the factors N and σ^2 these weights are simply the squares of the reciprocals of the standard errors of successive percentiles of a normal distribution. The proportional magnitudes, $2\pi z^2/4\,pq$ are the weights of Urban's Table. The factor $2\pi/4$ was chosen by Urban merely in order to make the maximum weight 1.

We may consider two cases in applying these weightings:

First: When neither σ nor $\bar{s}$ is known. In this case the sum of the squares of the residuals

$$\left(x_1 - \frac{s_1}{\sigma} + \frac{\bar{s}}{\sigma}\right), \left(x_2 - \frac{s_2}{\sigma} + \frac{\bar{s}}{\sigma}\right), \cdots \left(x_m - \frac{sm}{\sigma} + \frac{\bar{s}}{\sigma}\right)$$

is to be made a minimum, after each has been given its appropriate weight, w_1, w_2, $\cdots w_m$, as defined by equations [310].

$$w_1 = N_1 \frac{z^2_1}{p_1 q_1}, w_2 = N_2 \frac{z^2_2}{p_2 q_2}, \cdots w_m = N_m \frac{z^2_m}{p_m q_m} \quad \text{(Constant method weights)} \ldots\ldots [310]$$

The magnitudes z^2/pq are readily obtained, being the product of z/p and z/q of Table K–W. The magnitudes N are the numbers of cases in the successive experiments. By the usual method of least squares, the required values of $1/\sigma$ and $\bar{s}/\sigma$ are given by the solution of the two following simultaneous equations, in which Σ indicates a summation of m terms:

$$\Sigma wx - \frac{1}{\sigma}\Sigma ws + \frac{\bar{s}}{\sigma}\Sigma w = 0 \quad \ldots\ldots [311]$$

(Normal equations for threshold and dispersion calculations)

$$\Sigma wxs - \frac{1}{\sigma}\Sigma ws^2 + \frac{\bar{s}}{\sigma}\Sigma ws = 0 \quad \ldots\ldots [312]$$

Second: When the chief concern is with the determination of precision of judgment and $\bar{s}$ is known without experimental determination. Such situations may arise in the derivation of educational and psychological scales, such as drawing or composition scales where $\bar{s}$ is taken as equal to K. In this

case equation [311] only is necessary, as $\bar{s} = K$, a known quantity. Solving [311] for σ we have

$$\sigma = \frac{\Sigma ws - K \Sigma w}{\Sigma wx} \quad \text{(Calculation of dispersion when the threshold is known)} \ldots . [311\ a]$$

The following problem illustrates the steps involved in the method. The data are drawn from the educational field to show the value of psychophysical methods in a much wider field than that to which they are usually limited.

A judge is called upon to rank an English composition as better or worse than 40 standard compositions which are graded on a certain scale of merit. Ten of these forty have merit 38, eight have merit 50, six have merit 60, and 16 have merit 68. The rankings given by the judge and the calculation of the threshold and the dispersion are as follows:

TABLE LXXIII

Merit of Compositions Used as Standard	No.	Number of Times Sample Is Ranked Better than the Standard Used	Proportion of "Better" Judgments. p of Table K–W	x	$N\frac{z^2}{pq}$ i.e. w	wx	ws	wxs	ws^2
s									
38	10	7	.70	−.524401	5.757	−3.019	218.8	−114.7	8313
50	8	4	.50	.000000	5.093	.000	254.7	.0	12733
60	6	3	.50	.000000	3.820	.000	229.2	.0	13752
68	16	2	.875	1.150349	6.199	7.131	421.5	484.9	28664
					20.869	4.112	1124.2	370.2	63462
					Σw	Σwx	Σws	Σwxs	Σws^2

Thus the normal equations are:

$$4.112 - \frac{1}{\sigma} 1124.2 + \frac{s}{\sigma} 20.869 = 0$$

$$370.2 - \frac{1}{\sigma} 63462. + \frac{\bar{s}}{\sigma} 1124.2 = 0.$$

Their solution gives

$$\sigma = 19.55 \text{ and } \bar{s} = 50.02$$

We thus conclude that the integral of a normal distribution having a mean of 50.02 and a standard deviation of 19.55 is the best fit determination. If the purpose of the investigation has been to determine the merit of the sample, we conclude

that 50.02 is the best estimate of its true merit. The error of this value is unknown, but if the standards of comparison have been such that proportions, p, not greatly different from .5 have resulted, the standard error is probably in the neighborhood of $1.5\sigma/\sqrt{\Sigma N}$. If all the proportions are very large or very small, the error will be much larger than this. If it is known ahead of this calculation that the sample has a certain merit, let us say 45, then the calculation shows that the systematic error of the judge is 5.02 and that his chance error is represented by a distribution with standard deviation 19.55. Note that systematic error is synonymous with threshold, and standard error of judgment with the psychophysical measure of dispersion.

CHAPTER XIII

INDEX NUMBERS

Section 93. The Bearing of Purpose and Material upon Form of Index

The discussion in this chapter will be with reference to price ratios and averages of such ratios, as they are found to vary from time to time. The treatment does not, however, necessitate that price and time be the two variables. In dealing with size of certain organisms in a liquid media, length and temperature might be the two variates. Illustrations from other fields will be equally obvious.

In planning the construction of an index number in the field of economics three questions are important: (*a*) What is the purpose to be served by the proposed index? (*b*) What price and quantity data can be selected or collected to best serve this purpose? and, (*c*) What form of index is the best in the light of (*a*) and (*b*)?

(*a*) Though the chief treatment of this chapter is with (*c*) it should be borne in mind that differences in (*a*) and (*b*) can conceivably completely change the form of index which is most suitable. In particular a problem requiring an index, the meaning of which can be accurately grasped by a lay audience, cannot involve geometric and harmonic means; an index which, for the use that is to be made of it, must be reversible no matter what year is made the base, cannot be built upon quotations of commodities differing from year to year; an index which is required to serve the double purpose of being equally serviceable whether price relatives or quantity relatives are sought, cannot be asymmetrical with respect to prices and quantities; an index which is designed to picture an aggregate condition in an industry, country, or other unit,

cannot be based upon partial data unless it incorporates provision for estimation of omitted material; etc., etc.

Fisher (1921) has especially stressed the value of an aggregate index which is both a price and a trade index, permitting interpretation as to quantities involved as well as prices paid. He implies that an "unbiased" index meeting these conditions, of which there is more than one, is the index par excellence, answering all the essential problems. As to whether this is so is a question of economics and only secondarily of mathematics. For this reason the present treatment stresses this feature less than does Fisher. This does not imply a disagreement with Fisher but rather an indisposition to attempt to answer a problem which is in the main economic.

The number and nature of the commodities entering into an index depends upon the degree of accuracy required and the particular purpose to be served. They are consequent to the form of index used only because certain indexes require both price and quantity data while others are less exacting. Having determined the form of index, and knowing the purpose, and ruling out of consideration the index which is a complete survey of a field the question in choosing commodities is, what are the principles which should control in drawing a sampling? The fundamental principles of multiple correlation apply — high correlation with the purpose to be served and low intercorrelation. If a coal price index is being constructed from a small number drawn from a much larger number of quotations, the quotations should be chosen so that (*a*) each is as little correlated as possible with the other quotations included in the index, and (*b*) each is as highly correlated as possible with the other quotations in the field not included in the index. It is to be expected that commercial tendencies will conspire to prevent any quotation from markedly possessing both characteristics, in which case a balance must be struck between them: (*b*) is the more important if the number of quotations in the index is small, say not over six, but (*a*) is by far the more important if the number of quotations is large. In fact, quotations that are excellent for incorporation in an index number based upon a small number of items may be expected to be relatively inferior for incorporation in an index based

upon a large number of items. This brief observation as to the significance of correlation between commodity prices is, in the main, an addendum to, not in opposition to, the points involved in Mitchell's (1915) very thorough exposition of the question of what commodities should be included.

The preceding paragraphs merely touch upon the various phases of the problem of purpose and selection of material. No one source covers this adequately, but the reader will find a fairly complete treatment of all phases of the problem in the following selected list of references: Edgeworth (1896) and (1887, 88, 89, 90), Fisher (1913) and (1921), Knibbs (1912), Mitchell (1915), Pearson (1910, const.) and (1911, ops.), Walsh (1901) and (1921).

The succeeding treatment of topic (*c*) is taken with some modification and abridgment from Kelley (1921, cert.).

Section 94. The Meaning of a Price Ratio and of a Price Index

The price of a commodity in some one year, $p^1{}_1$ (the superscript designates the commodity, while the subscript designates the year), divided by the price of the same commodity in a second year, $p^1{}_2$, is $p^1{}_1/p^1{}_2$, and is called a price ratio. A composite of several such ratios purporting to portray a general relationship between prices in the two years is a price index, P_1/P_2. The fundamental concept in this is the ratio or geometric concept. Indices can be built upon many bases, but irrespective of the method of construction, the usual interpretation will involve this geometric concept. The lay reader will think that P_1 is a certain proportion of P_2, and P_2 is the inverse proportion of P_1. An index which is not reversible does not parallel the thought processes inherent in the concept "price ratio," and this more elementary concept, where reversibility is the rule, is the one by means of which "price index" is interpreted. Even writers who are quite aware that the index they are using is not reversible, use price ratios and price indices in such a way that it is obvious they expect the same sort of concept to be called up in the reader's mind; for example, "$p^1{}_1/p^1{}_2 = 122$, but $P_1/P_2 = 120$ so that, etc."

In so far as the concept P_1/P_2 is commonly of a different nature from $p^1{}_1/p^1{}_2$, it lies in the fact that P_1 and P_2 are averages, and $p^1{}_1$ and $p^1{}_2$ are single measures. Accordingly, to parallel customary thinking, P_1/P_2 should mean a reversible proportion between averages. What an "average" is may not be so definitely established in the minds of scientific people generally as is the idea "ratio," but probably the most common concept is that of arithmetic average or mean. We therefore have the somewhat anomolous situation of P_1/P_2 calling up the arithmetic concept when dealing with the two separate elements involved in it, but the geometric concept when dealing with the thing entire. Since this mixture of concepts seems likely to persist, the writer proposes as an important test of the excellence of an index number the closeness with which the operations involved in it parallel general thinking tendencies: *First and most important, reversibility of ratio, and second, arithmetic averages involved in the parts.*

Section 95. The Probable Errors of Various Indexes

That a price index has a probable error is a fact not always recognized and not entirely obvious, for it may easily happen that the price ratios are entirely reliable. It may be possible to say that the price of cotton at a certain time was $p^1{}_1$ and at a second time $p^1{}_2$. If the price quotations are accurate, then the price ratio $p^1{}_1/p^1{}_2$ is a true measure. The average of several such gives P_1/P_2, which is invariable. Therefore, P_1/P_2 has zero probable error as far as being the average of these particular things, but the very combining of them involves the assumption that the index has significance beyond the particular data from which it is calculated. The only exception would be when P_1 and P_2 are determined from all the possible data. As an example, let $p^1{}_1$ be the price of coal at a certain mine at the first date, $p^2{}_1$ the price at a second mine, . . . , $p^n{}_1$ the price at the last mine, and similarly for the p_2's. Then, since all the sources are involved, P_1/P_2 is the index of coal prices and has no probable error, except such as might be due to faulty quotations and calculations and could therefore, by proper care, be made negligible.

This is not the typical situation. Ordinarily but a few quotations are worked up into an index and the result taken as representative of an industry or a field. We therefore have quotations which are samplings of the prices in the industry, and the statistical methods for determining the reliability of samplings apply. The formulas for probable errors given in succeeding sections are based upon certain assumptions, including that of random sampling, but if 25 or more per cent of the possible quotations are utilized, material error in the formulas is introduced, the true probable errors being less than those given by the formulas. It is to be understood that by probable error in an index number is meant that which arises from incompleteness of data. In the following determinations of probable errors of index numbers as given by various formulas, the attempt is to see how closely one can approximate, by a sample, the number which would be obtained were all the possible data utilized in determining the same sort of index. The probable error indicates how closely the results from the sample may be expected to tally with the results from the whole. Should there be a constant tendency in the form of index used, systematically leading to too high or too low a value, we have a systematic error, which is entirely distinct and which is not measured by the size of the probable error.*

The reason why a few quotations can yield an index which is a close approximation to a general tendency is that there is a high correlation between the quotations included and those not included in the index but pertinent to the function being measured. If there are two hundred coal mines and quotations from a half dozen are taken, an index in close agreement with the true index based upon the two hundred may be expected, because of the high correlation between quotations at different mines. To say that there is a high correlation is not equivalent to saying that the prices at the different mines tend to approach the same level, but that they tend to main-

* In the tests of indices suggested in Section 97 there will be found none to the effect that an index should have no bias. The reason for this is that reversibility of ratio, or change of base, which is included as one of the tests, is not possible with a "biased" index. Fisher (1921) shows that an index may possess a bias due to form and a second bias due to base value weighting, and that these may exactly neutralize each other. Such a situation would, statistically, be the same as one not involving bias.

tain a uniform difference. Mine A, near tidewater, may sell at a certain price, p^1, much higher than that, p^2, at mine B, remote from a center of consumption, without indicating an economically abnormal condition in the coal trade. If p^1, p^2, and other similar measures are averaged, the probable error of this average is not given by the usual formula

$$\text{P. E.}_{\text{mean}} = .6745 \frac{\sigma}{\sqrt{N}}$$

due to the heterogeneity of, and to the correlation between, the p's. As an illustration, more extreme than mine quotations on coal, let us average the following prices:

Bacon per pound	$ .70
Bread per pound	.10
Potatoes per bushel	1.20
Apples per box	10.00
Average	$3.00
Standard deviation	4.06
P. E. (by above formula)	1.37

Now, presumably, the probable error of no single one of these quotations is as great as $1.37, and the average of them all will probably fluctuate but little. There probably is positive correlation between these food prices, a rise in one generally going with a rise in each of the others. These conditions are not those under which the probable error of an average is given by the usual formula. For statistical purposes there is much to be gained by having homogeneous uncorrelated material. We can secure measures which are nearly, if not entirely, homogeneous and uncorrelated by dealing with price ratios instead of prices.*

* In one sense, both prices and price ratios are very highly correlated, but these correlations have quite different statistical consequences. As the price of coal at mine A approaches p^1_1, due to correlation the price at mine B approaches what may be a very different value, p^2_1; but as the ratio, p^1_1/p^1_2, from the quotations of mine A approaches, as time changes, the value p, due to correlation, the ratio of the quotations from mine B may be expected to tend toward the same value p. (The rigorous proof of this statement would be necessary before the present treatment and statement of probable errors can be considered final. Whatever error is involved is of a conservative nature, as it almost certainly would tend to make the obtained probable errors too large.) Although correlation between prices tends to throw ratios together, it tends to keep prices apart. If, therefore, we deal with ratios, the effect of correlation has already operated upon the measures used, making the distribution of ratios more homogeneous, and as a consequence making the mean more reliable. In other words, the standard deviation of the ratios of prices at date 1 to those at date 2, σ_{12}, is reduced from what it would be were there no correlation between prices, so that by this very reduction, the probable error formula when applied to ratios takes account of the correlation between prices at two different dates. For a rigorous approach to the question of probable error of a ratio see Pearson (1910 const. and 1911 ops.).

Accordingly, if the price index showing prices in year 1 relative to year 2, called i_{12}, is given by the equation

$$i_{12} = \frac{P_1}{P_2} = \frac{1}{N}\Sigma\frac{p_1}{p_2} \quad \text{(Index formula 1)}. [313]$$

and if the standard deviation of the price ratios is σ_{12}, the probable error of i_{12} is given by

$$\text{P. E. } i_{12} = .6745\frac{\sigma_{12}}{\sqrt{N}}. \quad \text{(Probable error of index formula 1)}. [314]$$

Let us consider another kind of index,

$$i_{12} = \frac{P_1}{P_2} = \frac{\Sigma p_1}{\Sigma p_2}. \quad \text{(Index formula 2)}. . [315]$$

The complete probable error formula for this kind of index involves the correlation between the p's. (See Pearson, 1910, ops.) The index

$$i_{12} = \frac{1}{\Sigma w}\Sigma\left(\frac{p_1}{p_2}\right)w. \quad \text{(Index formula 3)}. [316]$$

will be more reliable than formula 1 if the weights, w, used are exactly or approximately proportionate to the values of the commodities involved. In general, the greater the price ratio the less the consumption and vice versa, so that the distribution of the weighted price ratios will have a smaller variability than the distribution of price ratios alone. If $w = p_2q_2$, the value of the transactions in year 2, the formula becomes

$$i_{12} = \frac{\Sigma p_1q_2}{\Sigma p_2q_2}. \quad \text{(Index formula 4)}. . . . . [317]$$

Formula 4 is but a type of formula 3. It is undoubtedly more reliable than either 1 or 2, but there are too many variables involved for the writer to attempt a calculation of its probable error based upon the data for two dates only. If, however, the commodities are divided into random halves and indexes determined from each half, the correlation between these sub-indexes may be calculated, and from it the probable error of the total index may be obtained, as follows:

Let there be n commodities, equally excellent as representa-

tive of the whole field, which are built up into the index i. In order to determine the probable error of i we may first build up two indexes, A and B, each based upon a random half of the commodities. Calculation of A and B for a number of dates will give two series, the correlation between which may be found. In doing this it is desirable that the time interval between successive indexes be sufficient to insure the relative independence of the commodity quotations involved. Just as the average of the prices of bread on January 1 of a certain year and on December 30 of the same year will in general give a truer average yearly price than the average of the prices on June 30 and July 1, because in the former case the two quotations are nearly independent while in the latter one has practically but a single quotation, so sub-indexes calculated at too short intervals of time scarcely constitute new data, but rather repetitions of old data. Were the correlation between successive quotations known, practical limits could be set giving periods shorter than which it would not be worth while to calculate sub-indexes. Having σ, the standard deviation of these sub-indexes, and having r, the correlation of the sub-indexes, we may determine the standard error of the average of the two sub-indexes, i.e., of the total index, i. As given by Kelley (1921, cert.), it is

$$\sigma_1 = \sigma \sqrt{\frac{1 - r}{2}}$$ (Standard error of an index in terms of the standard deviation and correlation of sub-indexes).....[318]

Note that r and σ must be obtained from the same series of sub-indexes.

The practical advantages of reporting two sub-indexes as well as the total index may well be as great as has been found to be the case in reporting two comparable measures in the fields of psychology and education. The probable error of any index may be determined if comparable sub-indexes are calculated and if the series of indexes covers a sufficient length of time to yield a reliable measure of correlation between sub-indexes. Probably 16 pairs of quarterly sub-indexes would suffice. Since a means of determining the standard error of any index is available, we may say that a second important

measure of the excellence of an index number is *the size of its probable error.**

Space will not permit a discussion of the probable errors of all the proposed types of indexes, but to point out the necessity of such discussions the writer has made an estimate, after more or less complete mathematical analysis of the relative size of the probable errors of the index numbers given in Table LXXIV, Section 97.

The one that seems the most reliable of all, and that also most completely meets other conditions except that of paralleling general thinking tendencies, is the weighted geometric mean index, in which the weights are roughly proportional to the reliabilities of the price ratios. This requirement as to weights is practically no limitation at all, as it is regularly approximated to by customary weighting devices. Practically without exception the observations of Mitchell (1915) as to what items to include in an index and what weights to give, are statistically equivalent to weighting price ratios according to reliability.

Section 96. The Accuracy and Flexibility of the Weighted Geometric Mean Index

The weights of the commodities involved in an index may be changed with much greater facility in the case of some indexes than of others. As soon as a commodity becomes archaic the proper thing to do is to withdraw it, and withdrawals and entrances are readily accomplished with the geometric index. The weighted geometric mean index formula is

$$i = \sqrt[\Sigma w]{\frac{(p^1{}_1)^{w_1}(p^2{}_1)^{w_2}\cdots(pn_1)^{w_n}}{(p^1{}_2)^{w_1}(p^2{}_2)^{w_2}\cdots(pn_2)^{w_n}}}. \quad \text{(Index formula 5)}\ldots[319]$$

* I judge from the limited abstract of his study that Fisher (1921) has calculated a large number of different indices from the same material and found that certain formulas give highly comparable results. The uniformity of indices involving the same data is not the problem of reliability here attacked. We are concerned with the problem of sampling. As to whether Professor Fisher has also compared an index determined from a part of his data with the same index as obtained from a larger part I cannot determine from the abstract, but if so it constitutes an experimental approach to the problem in hand. One would expect that the differences which Professor Fisher would find between an index based upon, let us say, $\frac{1}{4}$ of his data and one based upon the remaining $\frac{3}{4}$ would be somewhat larger than implied by the formula here given, as the index based upon the $\frac{3}{4}$ would be a fallible standard. A study of the uniformity of indices based upon the same data throws light upon the existence and the nature of systematic tendencies, or biases, but none whatever upon the error of sampling.

For convenience, and without any loss of generality, Σw may be made to equal 1. Thus, letting $\omega_1 = w_1/\Sigma w$, $\omega_2 = w_2/\Sigma w$, etc., and letting $\rho_1 = p^1{}_1/p^1{}_2$, $\rho_2 = p^2{}_1/p^2{}_2$, etc.,

$$i = \rho_1{}^{\omega_1}\rho_2{}^{\omega_2}\cdots\rho_n{}^{\omega_n}. \quad \text{(Index formula 5 } a\text{)} \ldots [319\ a]$$

Note that with this formula the index is reversible and that there is complete freedom in changing the base. Assuming as before that there is no correlation between ratios, the probable error is given by

$$\text{P. E.}_i = .6745\frac{i}{\Sigma w}\sqrt{\frac{w^2{}_1\sigma^2{}_1}{\rho^2{}_1} + \frac{w^2{}_2\sigma^2{}_2}{\rho^2{}_2} + \cdots \frac{w^2{}_n\sigma^2{}_n}{\rho^2{}_n}}$$

(Probable error of the weighted geometric mean index) . . [320]

in which the ρ's are successive price ratios and the σ's their standard deviations. As an approximation, the σ's may be considered to be equal to each other and to equal the standard deviation of the distribution of price ratios. In order that this probable error remain small, it is necessary that no one of the ratios w_1/ρ_1, w_2/ρ_2, etc., be exceptionally large.

$$\frac{w_1}{\rho_1} = \frac{w_1 p^1{}_2}{p^1{}_1}.$$

Letting $q^1{}_1$ equal the quantity of the commodity consumed, or in trade, it would be expected that $q^1{}_1 p^1{}_1$ would fluctuate much less than $p^1{}_1$, and whereas there might be danger of $p^1{}_1$ becoming extremely small or large there is not equal likelihood of $q^1{}_1 p^1{}_1$ doing so. Accordingly, if w_1 is approximately $= q^1{}_1 p^1{}_1$, then $w_1/\rho_1 = q^1{}_1 p^1{}_2$, a magnitude which is not likely to be extremely large. However, should a commodity change greatly in its relative importance, the weighting of it may easily be changed as follows:

Let it be desired to change the weight of the price ratio ρ_1 from w_1 to W_1, which we will say is a smaller weight. We need not impose the condition that $\rho_1 = i$. For $\rho_1 > i$ we will search the list of price ratios for (a) a ratio $> i$ which is underweighted, or (b) a ratio $< i$ which is overweighted. Suppose ρ_2 is such a ratio. Ordinarily there are a number of price ratios = 1.0, or i, or some other value which is the modal value. These may be combined and represented by ρ^s, where ρ is this modal value and s the sum of the weights of all the

ratios having this value ρ. Letting P stand for the product of all the terms other than ρ_1, ρ_2, and the ρ terms, we have

$$i = \sqrt[\Sigma w]{\rho^{w_1}\rho^{w_2}\rho^{s}P}$$

and it is desired to change this to

$$i = \sqrt[\Sigma W]{\rho^{W_1}\rho^{W_2}\rho^{S}P.}$$

The first index will equal the second in case

(1) $$w_1 + w_2 + s = W_1 + W_2 + S \quad \text{.....[321]}$$

and also

(2) $$\rho^{w_1}\rho^{w_2}\rho^{s} = \rho^{W_1}\rho^{W_2}\rho^{S}, \quad \text{.....[322]}$$

or, taking logarithms,

$$w_1 \log \rho_1 + w_2 \log \rho_2 + s \log \rho = W_1 \log \rho_1 + W_2 \log \rho_2 + S \log \rho \text{ .[323]}$$

W_1 is the new weight that has been assigned (this may be zero) so that everything involved is known except W_2 and S, and the solution of the two equations simultaneously will yield these. Ordinarily S will differ but slightly from s, and W_2 will differ from w_2 in the direction in which it is desirable it should differ. Thus, as a practical matter, the weight of any price ratio, whether equal to i or not, may be changed without affecting the index.

No other index, as far as the writer can determine, offers the extreme flexibility in changing weights, dropping or adding new items, here found to exist for the geometric mean index. Since this is so, the weights can be made such that extreme ratios are given small weights or eliminated. As a consequence, the probable error of such a weighted geometric mean index may be expected to be smaller than that of any other index mentioned. The excellence of this index seems to the writer so great as to warrant its use, even though it involves a change in the established habits of interpretation of the usual reader.

Section 97. Criteria for Judging of the Excellence of Indexes

Two criteria, the paralleling of habitual modes of thinking and reliability, have been proposed in judging the excellence of an index measure. Fisher (1913) has used eight other tests, three of them being tests only of "trade" indexes. It

would seem that these latter would be of particular importance only in case an index ceases to be a sampling and becomes an expression of the sum total of transactions involved. Table LXXIV, in part taken from Fisher (1913), gives "scores" of the most important index measures upon several tests or criteria of excellence.

Test 1: Reliability. In giving scores upon this point the writer has freely used his judgment in the case of indexes for which no simple probable error formula is available. More or less complete statistical analysis has preceded this scoring, but it is in no sense to be considered final. An "*s-i*" after a score means that no simpler way for calculating the probable error than by means of the correlation between comparable sub-indexes seems to be available. As the writer judges this test to be the most important of all, the scoring is 3, 2, 1, and 0, instead of 2, 1, and 0 — the larger the score, the higher the rating.

Test 2: Parallels habitual modes of thinking. Score 2, 1, 0.

The following tests are from Fisher.

Test 3: Proportionality. "A price index should agree with the price ratios if these all agree with each other." Stated algebraically:

$$\text{Given } \frac{p^1{}_1}{p^1{}_2} = \frac{p^2{}_1}{p^2{}_2} = \text{etc.} = i. \quad \text{Required that } \frac{P_1}{P_2} = i. \quad \ldots\ldots[324]$$

Score of 2 if true for any two years. Score of 1 if true only when year 2 is the base year.

Test 4: Entry and withdrawal. A price index should permit the entry and withdrawal of price ratios without changing the value of the index. Fisher uses a less general test: "A price index should be unaffected by the withdrawal or entry of a price ratio agreeing with the index." The scoring here follows Fisher, except for formula 5, which Fisher does not include in his list of 44, and for formulas 14 and 15 which are here scored higher than by Fisher.* Score 3, 2, 1, 0.

* Fisher scores both of these formulas zero on the basis of entrance and withdrawal of items. However, as shown by Kelley (1921 cert.) a new commodity, whose price ratio agrees with the index, may be introduced into index formula 15, without changing its value provided quantities are in the ratio,

$$\frac{q^1{}_1}{q^1{}_2} = \frac{ab(di - c)}{cd(a - bi)}$$

Test 5: Change of base. "The ratios between price indexes should be unaffected by reversing or shifting the base." Algebraically stated:

$$\text{Let } i_{12} = \frac{P_1}{P_2},\ i_{45} = \frac{P_4}{P_5}, \text{ etc. Required that } \frac{i_{24}}{i_{14}} = \frac{i_{32}}{i_{12}} = \frac{P_3}{P_1} = i_{31}. \quad [325]$$

Give score of 2 if true for any two years, score of 1 if only true when the base year and one other is involved, i.e., if only such equations as $\frac{i_{33}}{i_{13}} = i_{31}$, $\frac{i_{22}}{i_{42}} = i_{24}$, etc., hold.

Test 6: Change of unit of measurement. "The ratios between various price indexes should be unaffected by changing any unit of measurement." Score of 2 or 0.

Fisher has a "Determinateness" test which he describes in the words, "A price index should not be rendered zero, infinity, or indeterminate by an individual price becoming zero." This is but one phase of reliability and is therefore included in Test 1 above.

In the formulas listed the q's stand for quantities of commodities consumed or in trade and are weights of the p's. When weights not exactly equal to the q's are involved, the symbol w is used. It is of course assumed that care would be exercised in selecting these weights. p_0 and q_0 instead of p_2 and q_2 are used in those formulas in which the treatment of the data for the base year is unique. Test 5 is not completely met by any such formulas.

---

in which

$$a = \Sigma p_1 q_1$$
$$b = \Sigma p_2 q_1$$
$$c = \Sigma p_1 q_2$$
$$d = \Sigma p_2 q_2$$

$$i = \sqrt{\frac{\Sigma p_1 q_1}{\Sigma p_2 q_1} \times \frac{\Sigma p_1 q_2}{\Sigma p_2 q_2}}. \quad \text{(Index formula 15)}$$

Also, if quantities are in the ratio

$$\frac{q^1_1}{q^1_2} = \frac{\Sigma p_2 q_1}{\Sigma p_2 q_2}$$

a commodity whose price ratio is equal to the index may be introduced into index formula 14,

$$i_{12} = \frac{\frac{\Sigma p_1 q_1}{\Sigma p_2 q_1} + \frac{\Sigma p_1 q_2}{\Sigma p_2 q_2}}{2}. \quad \text{(Index formula 14)}$$

without changing its value.

TABLE

Scores of Index Numbers upon

Tests	(1) Type IA $\frac{1}{N}\Sigma\frac{p_1}{p_0}$ Carli Evelyn Economist Sauerbec, Soether	(2) Type IA $\frac{\Sigma\frac{p_1}{p_0}w}{\Sigma w}$ Young Falkner Dun
1 Reliability,—Smallness of P. E.	.5 s-i	1.5 s-i
2 Parallels habitual mode of thinking . . .	1.	1.
3 Proportionality	2.	2.
4 Entry and withdrawal	2.	2.
5 Change of base	.0	.0
6 Change of unit of measurement	2.	2.
Totals	7.5	8.5

TABLE

Scores of Index Numbers upon

Tests	(8) Type IV $\frac{\Sigma p_1}{\Sigma p_2}$ Bradstreet	(9) Type IV $\frac{\Sigma p_1 w}{\Sigma p_2 w}$ Lowe	(10) Type IV $\frac{\Sigma p\,\frac{q_0+q_1}{2}}{\Sigma p_0\,\frac{q_0+q_1}{2}}$ Edgeworth Marshall	(11) Type IV $\frac{\Sigma p_1\sqrt{q_0q_1}}{\Sigma p_0\sqrt{q_0q_1}}$ Scrope and Walsh	(12) Type IV Also Type IA $\frac{\Sigma p_1q_0}{\Sigma p_0q_0}$ Scrope Sidgwick Sauerbeck Giffen
1	1.5 s-i	2.5 s-i	2.5 s-i	2.5 s-i	2. s-i
2	2.	2.	1.5	1.5	1.
3	2.	2.	1.	1.	2.
4	2.	2.	1.	1.	2.
5	2.	2.	1.	1.	.0
6	.0	2.—	2.	2.	2.
Totals . .	9.5	12.5 —	9.0	9.0	9.0

Type IA: Arithmetic average of ratios
Type IH: Harmonic average of ratios
Type II: Median of ratios
Type III: Geometric average

Formulas 7 and 9, which are given the highest scores, involve weights, w, instead of quantities, q. There is great flexibility in each of these so that if a weight is adopted, let us say in the first instance upon the basis of quantities (if using formula 9) or values (if using formula 7) in trade, which tends to become

LXXIV

Basis of Six Tests of Excellence

(3) Type IA $\frac{\Sigma \frac{p_1}{p_0} p_1q_1}{\Sigma p_1q_1}$ Palgrave	(4) Type II Median Value of $\frac{p^1_1}{p^1_0}, \frac{p^2_1}{p^2_0} \ldots$ Edgeworth	(5) Type II Weighted median	(6) Type III Also Type V $\frac{\sqrt[n]{p^1_1 p^2_1 \ldots}}{\sqrt[n]{p^1_2 n^2_2 \ldots}}$ Jevons Westergaard	(7) Type III Also Type V Weighted geom. mean	Tests
1. s-i	2.	3.	1.	3.	1
1.	1.	1.	.5	.5	2
1.	2.	2.	2.	2.	3
1.	1.—	1.—	2.	3.	4
.0	1.—	1.—	2.	2.	5
2.	2.	2.	2.	2.	6
6.0	9.0 —	10.0 —	9.5	12.5	Totals

LXXIV—*Continued*

Basis of Six Tests of Excellence

(13) Type IV Also Type IH $\frac{\Sigma p_1q_1}{\Sigma p_0q_1}$ Scrope Sidgwick Sauerbeck Giffen	(14) Type VI Arith. average of (12) and (13) Sidgwick Drobisch	(15) Type VI Geom. average of (12) and (13)	(16) Type V $\frac{\Sigma p_1q_1}{\Sigma q_1} \frac{\Sigma p_2q_2}{\Sigma q_2}$ Drodisch Rawson-Rawson	(17) Type V $\frac{\Sigma p_1q_1}{\Sigma p_2q_2} \frac{\sqrt[n]{q^1_1q^2_1 \ldots}}{\sqrt[n]{q^1_2q^2_2 \ldots}}$ Nicholson Walsh	Tests
2. s-i	2.5 s-i	2.5 s-i	2. s-i	2. s-i	1
.5	.5	1.5	.5	.5	2
1.	1.	1.	.0	.0	3
1.	1.	1.	.0	.0	4
.0	.0	1.	2.	2.	5
2.	2.	2.	.0	2.	6
6.5	7.0	9.0	2.5	4.5	Totals

Type IV: Quotient of aggregates
Type V: Quotients of functions of data of single years
Type VI: Composites of preceding types

unreasonable, it can be changed without affecting the index between the year when the change is made and the preceding year. If years from early to late are designated by 1, 2, 3, 4 and if a formula-7 index number is started at the end of the first year, using weights proportionate to the values of the

commodities in trade, and continues until the beginning of year 4 before a change in weights is desirable, a change can at that time be made which will preserve the index i_{34} and its reciprocal i_{43}. The new weighting would probably give an i_{42} and an i_{41}, were they to be calculated, which would be slightly different from those given by the equations:

$$i_{42} = \frac{i_{43}}{i_{23}} \text{ and } i_{41} = \frac{i_{43}}{i_{13}}$$

which would exactly hold had no change in weights been made. This difference will usually be small, but if an index is demanded permitting changes in weightings and at the same time enabling the use, with exactness, of any year as base, it may be made by the expenditure of a little more labor.

Section 98. The Use of Any Year as Base

Formula 12 (or 13) in which there are no parameters, or flexible weightings, will serve as a foundation:

$$i_{12} = \frac{\Sigma p_1 q_2}{\Sigma p_2 q_2}.$$

Let M_1 = the mean of the p_1's
m_1 = the mean of the q_1's
S_1 = the standard deviation of the p_1's
s_1 = the standard deviation of the q_1's
r_{11} = the correlation between the p_1's (represented by the first subscript) and the q_1's (represented by the second subscript).

Symbols with other subscripts have comparable meanings, e.g., r_{24} = the correlation between the p_2's and the q_4's. Then,

$$\left.\begin{aligned} \Sigma\, p_1 q_2 &= N\,(M_1 m_2 + r_{12} S_1\, s_2) \\ \Sigma\, p_2 q_2 &= N\,(M_2 m_2 + r_{22} S_2\, s_2) \end{aligned}\right\} \quad \ldots\ldots\ldots\ldots\ldots [326]$$

Consequently, the numerator and the denominator for the index between any two years may be built up if the means, standard deviations, and correlations are known. The data required may be calculated each year, as the data for the

year become available, and tabulated in such a table as the following:

Data for Determining Index with Any Desired Year as Base

Years		Mp	Mq	S	s	r_{pq}: p for Year Indicated in Stub and q for Number of Years Indicated Earlier (−) or Later (+)											
						−32	−16	−8	−4	−2	−1	0	+1	+2	+4	+8	+16	+32
1919	9	+		+						+								
1918	8																	
1917	7	*	+×*	*	+×*							*						
1916	6																	
1915	5																	
1914	4																	
1913	3																	
1912	2																	
1911	1	×		×											×			

If it is desired to make 1917 the base and to express the prices in 1919 and 1911 relative to it, then $\Sigma p_9 q_7$ is determined from the magnitudes recorded in the compartments in which there is "+"; $\Sigma p_1 q_7$ from the compartments in which there is "×"; and $\Sigma p_7 q_7$ from the compartments in which there is "*."

The table as drawn up does not provide space for all the possible correlation coefficients. With such care as could be taken in choosing the units of quantity, the correlation coefficients could be made to vary from year to year in a very regular manner, thus enabling interpolation with high accuracy. There is complete freedom in changing the weights of commodities, but it should be noted that a commodity "dropped" continues as one of zero price and zero quantity — in other words, the N has not been decreased by "dropping" the commodity. To change the weight of a commodity price from w to w' demands a warrant. Let us say that such warrant is found in the ratio of the quantities consumed. No less warrant is necessary when w' is zero. An article once included in the index should come out only in case it becomes practically obsolescent. No distortion of any index would result in this case. We may of course take out a commodity under other conditions without affecting some one particular index.

APPENDIX A

LIST OF IMPORTANT SYMBOLS

When dealing with a single variable:

1. N designates the total population.
2. n is used as an exponent or subscript, or as the population of a sub-sampling.
3. X designates a gross score, i.e., a score as a deviation from zero in the quantity scale being considered.
4. x designates a score as a deviation from the mean.
5. ξ designates a score as a deviation from an arbitrary origin.
6. M designates the arithmetic mean.
7. Mdn designates the median ($= P_{.50}$).
8. Mo designates the mode.
9. p designates the proportion of cases lying below the $100\,p$ percentile, — to the left of a dichotomic point in a frequency polygon.
10. P_p designates the value of the $100\,p$ percentile.
11. q is defined by $p + q = 1$.
12. U.Q. designates the upper quartile ($= P_{.75}$).
13. L.Q. designates the lower quartile ($= P_{.25}$).
14. Q designates the quartile deviation, or semi-interquartile range ($=$ [U.Q. $-$ L.Q.]/2).
15. D designates the 10–90 percentile range ($= P_{.90} - P_{.10}$).
16. A.D. designates the average deviation, i.e., the mean deviation from the mean.
17. σ designates the standard deviation from the mean of scores in a distribution.
18. P.E. designates the probable error ($= .6744898\sigma$)
19. s designates the standard deviation from some point other than the mean.

20. Σ designates a summation of scores of the sort indicated.
21. S designates a summation of summations, or of elements other than individual scores.
22. σ with a subscript designates the standard error of the constant represented by the subscript.
23. P.E. with a subscript designates the probable error of the constant represented by the subscript.
24. i designates the class interval, or width of base of a given class in a frequency polygon.
25. v designates the value of the lower boundary of a class interval.
26. v' designates the value of the upper boundary of a class interval.
27. f designates the frequency in a class interval.
28. F designates the sum of the frequencies below a given class interval.
29. F' designates the sum of the frequencies above a given class interval.
30. Δ or δ designates the difference between the mean and arbitrary origin (= M-Arb. orig. = $\bar{\mu}_1$).
31. $\mu_1, \mu_2, \cdots \mu_n$ designate the moments from the mean (*a*) without application of Sheppard's corrections if they are inconsequential for the problem in hand, or (*b*) after application of Sheppard's corrections if they are used.
32. $\nu_1, \nu_2, \cdots \nu_n$ designate the moments from the mean before application of Sheppard's corrections in problems in which Sheppard's corrections are used.
33. $\bar{\mu}_1, \bar{\mu}_2, \cdots \bar{\mu}_n$ or $\bar{\nu}_1, \bar{\nu}_2, \cdots \bar{\nu}_n$ designate moments from an arbitrary origin.

When dealing with the normal distribution:

A normal distribution in which $N = 1$ and $\sigma = 1$ will be referred to as a "unit normal distribution."

In the general normal distribution, x as defined in 4, σ as defined in 17:

34. y designates the ordinate per unit interval (= zN/σ [as defined in 36 and 17]).

In the "unit normal distribution":

35. x designates a deviation from the mean
$$\left(= \frac{x \text{ [as defined in 4]}}{\sigma \text{ [as defined in 17]}}\right).$$

36. z designates the ordinate
$$\left(= \frac{y\,\sigma \text{ [as defined in 13 and 17]}}{N}\right).$$
p and q as defined in 9 and 11 ($p = \int_{-\infty}^{x} z\,dx = \frac{1}{2}[1 - \alpha]$ of Sheppard).

37. Corresponding to a deviation x_1 we have p_1, q_1, z_1; or corresponding to a proportion p_1, we have q_1, z_1, and x_1.

38. I designates $\int_{0}^{x} z\,dx$ ($=\alpha/2$ of Sheppard).

When dealing with unimodal distributions:

39. y_0 designates the ordinate at the origin (generally at the mean, the mode or a boundary).

40. m is an exponent. If two exponents are needed, m_1 and m_2 are used.

41. a in general designates the distance between the origin and a finite boundary. If two boundaries are finite, a_1 and a_2 are used.

When dealing with price indexes:

42. $p^t{}_s$ designates the price of commodity t at date s.

43. $q^t{}_s$ designates the amount consumed, or in trade, of commodity t at date s.
If few commodities are involved, subscripts are arabic numbers, and superscripts are primes.

44. p_s designates the price of an unspecified commodity at date s.

45. q_s designates the quantity consumed, or in trade, of an unspecified commodity at date s.

46. ρ_{su} designates a price ratio or the ratio of the price at date s to the price at date u $\left(= \frac{p_s}{p_u}\right)$.

47. P_s is a composite, weighted or otherwise, of the prices of several commodities at date s.

48. i_{su} designates a price index or the ratio of a composite of prices at date s to prices at date u $\left(= \frac{P_s}{P_u}\right)$.

49. w designates the weight given to a price, p, when this weight differs from q.

When dealing with correlated series:

50. Symbols as given in 3, 4, 5. Corresponding symbols for the second series are Y, y, ζ.

51. A second notation utilizes symbols 1, 3, 4, 5, 6, 7, 8, 9, 11, 15, 16, 17, 18 and 19, with subscript 1 added to represent the first variable, and a subscript 2 added to represent a second variable.

52. $\sigma_x = \sigma_1$, and $\sigma_y = \sigma_2$.

53. Δ has the meaning as in 28, with reference to the first variable, and δ this meaning with reference to the second variable.

54. z_1 designates the first variable expressed as a standard measure $-(= x_1/\sigma_1)$. z_2 designates the second variable expressed as a standard measure $-(= x_2/\sigma_2)$. See also 36.

55. r designates the product moment correlation coefficient between two series.

56. r is also used where specially noted, to designate bi-serial r, Sheppard's cos $2\pi\beta$ correlation, and occasionally other specially designated correlation coefficients.

57. ρ designates the correlation coefficient, based upon the squares of differences in rank.

58. R designates Spearman's foot-rule correlation coefficient. See also 86.

59. r_t designates the tetrachoric correlation coefficient.

60. $\sigma_{1.2}$ designates the mean standard deviation of the x-arrays from the regression line, i.e., it is the standard error of estimate of variable 1, knowing variable 2.

61. $\sigma_{2.1}$ designates the standard error of estimate of 2, knowing variable 1.

62. σ_a designates the mean standard deviation of the x-arrays from the means of the arrays.

63. x designates the value of x as estimated from a knowledge of y by means of the regression equation.

64. $\overline{X}$ designates the value of X as estimated from a knowledge of Y by means of the regression equation.

65. $\overline{y}$ and $\overline{Y}$ have comparable meanings to 63 and 64 interchanging the variables.

66. In general, a symbol with a superior bar stands for an estimated value of a variable, or for an average, but note 33, 81 and 82.

67. b_{12} designates the regression of the x's upon the y's, or the slope of the regression line used in estimating x's, knowing y's.

68. b_{21} designates the regression of the y's upon the x's.

69. h designates the grouping interval for the first variable ($= i$, in 24), and k, the grouping interval for the second variable.

70. χ is the first variate when no grouping is resorted to. It is not related to χ^2, of 99.

71. γ is the second variate when grouping is not resorted to. It is not related to γ as found in the equations of certain curves.

72. r, η, and C with a subscript preceding, such as subscript ${}_m r$, ${}_f \eta$, etc., designate a coefficient, after some correction has been made.

73. r with ∞ as one of the subscripts designates a correlation with a true score, i.e., a correlation corrected for attenuation.

74. $r_{\infty\infty}$ designates the correlation between two true scores, i.e., the correlation corrected for the attenuation in case of both variables.

75. k designates the coefficient of alienation or the proportionate improvement in estimate, due to the existence of correlation ($= \sqrt{1 - r^2}$). See also 85.

76. p with two subscripts designates a product moment. Distinguish between this and 9, 89 and 117.

77. d designates a difference between two scores. These scores may be rank positions.

78. η_{12} is the correlation ratio of x upon y, and η_{21} that of y upon x.

79. ζ is the test for linearity $(= \eta^2 - r^2)$. Distinguish between this and 50.

80. w represents an arbitrary weight. See also 91.

81. r_i designates the average inter-correlation between a number of independent variables.

82. r_c designates the average correlation between a criterion and a number of variables.

83. o used as a subscript designates the criterion.

84. Σ designates the standard deviation of scores in a second range when the standard deviation in the first range is σ. Distinguish between this and 20.

85. K designates the alienation coefficient in a second range, when the alienation coefficient in the first range is k. Distinguish between 85, 87 and 88.

86. R designates the correlation coefficient in a second range, when the correlation coefficient in the first range is r. Note also 58.

87. K^2 designates the mean of a summation. See formula [205].

88. κ designates the number of categories in a quantitative or qualitative distribution. λ designates the number of categories in a second quantitative or qualitative distribution.

89. p is the greater of two proportions which total 1.0, in a correlation table. See also 9.

90. α, β, γ, δ are the proportions in the four cells of a four-fold correlation table.

91. v and w with subscripts designate certain tetrachoric correlation functions. Distinguish between 25, 26, 80 and 91.

92. ϕ designates product-moment correlation between two two-point distributions. This is Pearson's r_{hk}, and also Yule's theoretical value of r.

93. ϕ^2 designates the mean square contingency. In the case of a four-fold only, it equals ϕ of 92 squared.

94. Q designates Yule's coefficient of association. Distinguish between 14 and 94.

95. ω designates Yule's coefficient of colligation.

96. $m_{ss'}$ designates the theoretical cell frequency.

97. $n_{ss'}$ designates the observed cell frequency.

98. $d_{ss'}$ designates the cell divergence $(= n_{ss'} - m_{ss'})$.

99. χ^2 designates the square contingency. See also 70.

100. P designates the probability of a divergence as great or greater than that obtained, arising as a matter of chance.

101. σ_{kk} designates the standard error of the k'th difference correlation coefficient.

When dealing with three or more correlated variables:

102. $x_{0.12\ldots n}$ designates the residual in the criterion, or error of estimate of the criterion, after regression equation estimation of it by means of the other variables $(= x_0 - \bar{x}_0)$.

103. x_0 designates the value of the criterion as estimated from the other variables.

104. $z_0 = x_0/\sigma_0$; $\bar{z}_0 = \bar{x}_0/\sigma_0$; $z_{0.12\ldots n} = x_{0.12\ldots n}/\sigma_0$; etc.

105. $r_{0.12\ldots n}$ designates the multiple correlation coefficient between the criterion and the regression equation combination of the independent variables.

106. $k_{0.12\ldots n}$ designates the multiple alienation coefficient between the criterion and the regression equation combination of the independent variables.

107. $\sigma_{0.12\ldots n}$ designates the standard error of estimate of the criterion, when estimated by means of the regression equation.

108. $r_{01.23\ldots n}$ designates the partial correlation coefficient between the criterion and variable 1, the other variables being constant.

109. $k_{01.23\ldots n}$ designates the partial alienation coefficient between the criterion and variable 1, the other variables being constant.

110. $\beta_{01\cdot 23\ldots n}$ designates the partial regression of the criterion upon variable 1, the other variables being constant, not allowing for unequal standard deviations of the variables.

111. $b_{01\cdot 23\ldots n}$ designates the partial regression of the criterion upon variable 1, the other variables being constant, taking into account the standard deviations of the variables ($= \beta_{01\cdot 23\ldots n}\, \sigma_0/\sigma_1$).

112. Δ designates the major determinant.

113. Δ_{pq} designates the determinant obtained by taking out the p'th row and the q'th column from the major determinant.

114. c designates the weighted composite of scores, generally slightly different from the regression equation composite.

115. u designates the one variable in the c composite which is treated in a unique manner.

116. $c - u$ designates the c composite, after deduction of the variable treated uniquely.

117. p designates any one of the variables, other than u, in the c composite.

118. D designates the distance from the stump to the mean of a complete normal distribution, in case of truncation ($= x\sigma$). See also 15.

119. σ_{-} designates the standard deviation of a weighted average.

THE GREEK ALPHABET

A	α	Alpha	I	ι	Iota	P	ρ	Rho
B	β	Beta	K	κ	Kappa	Σ	σ	Sigma
Γ	γ	Gamma	Λ	λ	Lamba	T	τ	Tau
Δ	δ	Delta	M	μ	Mu	Υ	υ	Upsilon
E	ϵ	Epsilon	N	ν	Nu	Φ	ϕ	Phi
Z	ζ	Zeta	Ξ	ξ	Xi	X	χ	Chi
H	η	Eta	O	o	Omicron	Ψ	ψ	Psi
Θ	θ	Theta	Π	π	Pi	Ω	ω	Omega

APPENDIX B

BIBLIOGRAPHY

Arranged chronologically under authors

ANDERSON, VON O.: Nochmals über "The elimination of spurious correlation due to position in time or space"; *Biom.*, Vol. X, 1914.

ANGELL, FRANK: "On judgments of like"; *Am. Jour. Psyc.*, Vol. XVIII, 1907.

BARLOW'S *Tables of squares, cubes, square-roots, cube-roots, and reciprocals of all integer numbers up to 10,000;* E. and F. N. Spon, Lond. and New York.

BELL, JULIA: "Tables to facilitate calculation of the rhinal indices"; *Biom.*, Vol. VIII, 1912.

BERGSTROM, SVERKER: "Sur les moments de la fonction de corrélation normale de *n* variables"; *Biom.*, Vol. XII, 1918.

BERNOULLI, J.: "Ars conjectandi, opus posthumum: Accedit tractatus de seriebus infinitis, et epistola gallicè scripta de ludo pilae reticularis," 1713. (A German translation in Ostwald's *Klassiker der exakten Wissenschaften*, Nos. 107, 108.)

BERTRAND, J. L. F.: *Calcul des probabilités.* Gauthier-Villars, Paris, 1889.

BLAKEMAN, JOHN: "On tests for linearity of regression in frequency distributions"; *Biom.*, Vol. IV, 1905.

BLAKEMAN, J. and PEARSON, KARL: "On the probable error of the coefficient of mean square contingency"; *Biom.*, Vol. V, 1906.

BOOLE, G.: *Laws of Thought*, 1854.

BOREL, É.: *Éléments de la théorie des probabilités.* Hermann, Paris, 1909.

BORING, EDWIN G.: "Mathematical *vs.* scientific significance"; *Psyc., Bul.*, Vol. XVI, No. 10, 1919.

——: "The logic of the normal law of error in mental measurement"; *Am. Jour. of Psyc.*, Vol. XXXI, 1920.

BOWLEY, A. L.: "Relation between the accuracy of an average and that of its constituent parts"; *Jour. Roy. Stat. Soc.*, Vol. LX, p. 855, 1897.

——: *Elements of statistics*, third edition; P. S. King, London, 1907.

BRAVAIS, A.: "Sur les probabilités des erreurs de situation d'un point," *Memoires . . . l'Academie royale des sciences de l'institute de France; sciences mathematique et physique*, Vol. IX, pp. 255–332, 1846.

BRINTON, WILLARD COPE: *Graphic Methods for Presenting Facts, N. Y. Engineering Mag.*, 1914.

BROWN, WILLIAM: *The Essentials of Mental Measurement;* Cambridge University Press, Lond., 1911.

BROWN, WILLIAM and THOMSON, GODFREY H.: *Essentials of Mental Measurement;* Cambridge, 1921.

BRUNT, DAVID: *Combination of Observations;* Putnam, 1917.

CARVER, HARRY C.: "Mathematical representation of the frequency distribution"; *Quar. Am. Stat. Assn.*, Vol. XVII, 1921.

CAVE, BEATRICE M. and PEARSON, KARL: "Numerical illustrations of the variate difference correlation method"; *Biom.*, Vol. X, 1914.

CHARLIER, C. V. L.: "Ueber das Fehlergesetz"; *Arkiv för Matematik*, Vol. II, Stockholm, 1905.

——: "Researches into the theory of probability"; *Meddelanden från Lunds Astronomiska Observatorium*, Lunds Universitets Årsskrift. N. F. Afd. 2, Bd. 1, N.: r. 5, 1906.

COTSWORTH, M. B.: *The Direct Calculator*, Series O. (Product table to 1000 × 1000.) M'Corquodale and Co., London.

COURNOT, A. A.: *Exposition de la théorie des chances et des probabilités*, 1843.

CRELLE, A. L.: *Rechentafeln.* (Multiplication table giving all products up to 1000 × 1000.) G. Reimer, Berlin.

CZUBER, E.: *Wahrscheinlichkeitsrechnung und ihre Anwendung auf Fehlerausgleichung, Statistik und Lebensversicherung;* Teubner, Leipzig, Ed. 1, 1903; Ed. 2, 1908–10; Ed. 3, 1921.

DAVENPORT, C. B.: *Statistical Methods;* John Wiley and Sons, New York and Chapman and Hall, Lond., 1904.

DAY, EDMUND E.: "Classification of statistical series"; *Quar. Am. Stat. Assn.*, New Series, No. 128, Vol. XVI, 1919.

——: "Standardization of the construction of statistical tables"; *Quart. Am. Stat. Assn.*, New Series, No. 129, Vol. XVII, March, 1920.

DEMORGAN, A.: "Treatise on the theory of probabilities" (*Encyclopedia Metropolitana*); 1837.

DICKSON, J. D. HAMILTON: Appendix to (Galton, 1886), *Proc. Roy. Soc.*, Vol. XI, p. 63, 1886.

DODD, EDWARD L.: *Error-risk of the median compared with that of the arithmetic mean;* Bul. 323, University of Texas, 1914.

DOODSON, ARTHUR T.: "Relation of the mode, median and mean in frequency curves"; *Biom.*, Vol. XI, 1917.

DUFFELL, J. H.: "Tables of the Γ function"; *Biom.*, Vol. VII, 1909.

EDGEWORTH, F. Y.: "Observations and statistics: An essay on the theory of errors of observation and the first principles of statistics"; *Cambridge Phil. Trans.*, Vol. XIV, p. 139, 1885.

——: "Problems in probabilities"; *Phil. Mag.*, 5th Series, Vol. XXII, p. 371; 1886.

——: "The choice of means"; *Phil. Mag.*, 5th Series, Vol. XXIV, p. 268; 1887.

——: Reports of the committee appointed for the purpose of investigating the best methods of ascertaining and measuring variations in

the value of the monetary standard; *British Association Reports* (p. 247), 1887; (p. 181), 1888; (p. 133), 1889; and (p. 485), 1890.

——: "On correlated averages"; *Phil. Mag.*, 5th series, Vol. XXXIV, p. 190; 1892.

——: Article "Index numbers" in *Palgrave's Dictionary of Political Economy*, Vol. II; Macmillan, 1896.

——: "On the representation of statistics by mathematical formulae"; *Jour. Roy. Stat. Soc.*, Vol. LXI, 1898; Vol. LXII, 1899; and Vol. LXIII, 1900.

——: "The law of error"; *Camb. Phil. Trans.*, Vol. XX, pp. 36–65 and 113–141; 1904.

——: "The generalized law of error, or law of great numbers"; *Jour. Roy. Stat. Soc.*, Vol. LXIX, p. 497; 1906.

——: "On the representation of statistical frequency by a curve"; *Jour. Roy. Stat. Soc.*, Vol. LXX, p. 102; 1907.

——: On the probable errors of frequency constants; *Jour. Roy. Stat. Soc.*, Vol. LXXI, pp. 381, 499, 651; 1908. Addendum, Vol. LXII, p. 81; 1909.

——: "Index numbers," *Dictionary of Political Economy*, Vol. II, pp. 384–387, Macmillan, 1910.

ELDERTON, W. PALIN: "Tables for testing the goodness of fit of theory to observation"; *Biom.*, Vol. I, 1902.

——: "Interpolation by finite differences"; *Biom.*, Vol. II, 1902.

——: "Tables of powers of natural numbers and of the sums of powers of the natural numbers from 1–100"; *Biom.*, Vol. II, 1903.

——: Notes on statistical processes; "An alternative method of calculating the rough moments from the actual statistics," "The application of certain quadrature formulae," "Adjustment of moments"; *Biom.*, Vol. IV, 1905.

——: *Frequency Curves and Correlation;* Layton, London, 1906.

——: "Some notes on interpolation in *n*-dimension space"; *Biom.*, Vol. 6, 1908.

EVERETT, J. D.: "On a new interpolation formula," *Jour. Inst. Actuaries*, Vol. XXV, 1901.

EVERITT, P. F.: "Tables of the tetrachoric functions for fourfold correlation tables"; *Biom.*, Vol. VII, 1910.

——: "Supplementary tables for finding the correlation coefficient from tetrachoric groupings"; *Biom.*, Vol. 8, 1912.

——: "Quadrature coefficients for Sheppard's formula (*c*)"; *Biom.*, Vol. XII, 1919.

FECHNER, G. T.: *Kollektivmasslehre*, herausgegeben von G. F. Lipps; Englemann, Leipzig, 1897.

FILON, L. N. G. and PEARSON, KARL. "On the probable errors of frequency constants and on the influence of random selection on variation and correlation"; *Phil. Trans.*, Vol. 191, pp. 229–311, 1898.

FISHER, ARNE: *Mathematical Theory of Probabilities;* Macmillan, 1915. Second Edition, enlarged, Macmillan, 1922.

FISHER, IRVING: *Purchasing Power of Money;* Rev. ed., Macmillan, 1913.

——: "Best form of index number"; *Quar. Am. Stat. Assn.*, March, 1921.

FISHER, R. A.: "On an Absolute Criterion for fitting frequency curves"; *Messenger of Math.*, New Series, Vol. XLI, pp. 155–160, Cambridge, 1912.

——: "On the distribution of the standard deviations of small samples"; *Biom.*, Vol. X, 1915.

——: "Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population"; *Biom.* Vol. X, 1915.

FOUNTAIN, H.: "Memorandum on the construction of index-numbers of prices"; in the *Board of Trade report on wholesale and retail prices in the United Kingdom*, 1903.

GALTON, FRANCIS: "Family likeness in stature"; *Proc. Roy. Soc.*, Vol. XL, p. 42; 1886.

——: "Correlations and their measurement"; *Proc. Roy. Soc.*, Vol. XLV, pp. 136–145, 1888.

——: *Natural Inheritance;* Macmillan and Co., 1889.

——: "Grades and deviates"; *Biom.*, Vol. V, 1907.

GARNETT, J. C. MAXWELL: *Education and World Citizenship.* (With a statistical appendix), Cambridge University Press, 1921.

GAUSS, C. F.: Méthode des moindres carrés: Mémoires sur la combinaison des observations, traduits par J. Bertrand, 1855.

GIBSON, WINIFRED: "Tables for facilitating the computation of probable errors"; *Biom.*, Vol. IV, 1906.

GREENWOOD, M.: "On errors of random sampling in certain cases not suitable for the application of a 'normal' curve of frequency"; *Biom.*, Vol. IX, 1913.

GROVE, C. C.: "Mathematics and psychology"; *Math. Teacher*, Vol. IX, 1916.

HARRIS, J. ARTHUR: "On the calculation of intra-class and inter-class coefficients of correlation from class moments when the number of possible combinations is large"; *Biom.*, Vol. IX, 1913.

——: "On spurious values of intra-class correlation coefficients arising from disorderly differentiation within the classes"; *Biom.*, Vol. X., 1914.

——: "A contribution to the problem of homotyposis"; *Biom.*, Vol. XI, 1916.

HASKELL, ALLEN C.: *How to Make and Use Graphic Charts;* Codex Book Co., New York, 1919.

HERON, DAVID: "An abac to determine the probable errors of correlation coefficients"; *Biom.*, Vol. VII, 1910.

——: "On the probable error of a partial correlation coefficient"; *Biom.*, Vol. VII, 1910.

——: "The danger of certain formulae suggested as substitutes for the correlation coefficient"; *Biom.*, Vol. VIII, 1911.

HOLZINGER, KARL J.: Communication "On the assumption that errors of estimate are equal in narrow and wide ranges"; *Jour. of Ed. Research*, Vol. IV, No. 3, p. 237, 1921.

HOOKER, R. H. "The correlation of the weather and the crops"; *Jour. Roy. Stat. Soc.*, Vol. LXX, 1907.

HOOKER, R. H. and YULE, G. U.: "Note on estimating the relative influence of two variables upon a third"; *Jour. Roy. Stat. Soc.*, Vol. LXIX, p. 197, 1906.

HUNTINGTON, EDWARD V.: *Handbook of Mathematics for Engineers*, New York, 1918.

——: "Mathematics and statistics, with an elementary account of the correlation coefficient and the correlation ratio"; *Am. Math. Mo.*, Vol. XXVI, pp. 421–435. Dec., 1919.

ISSERLIS, L.: "On the partial correlation ratio. Part I. Theoretical"; *Biom.*, Vol. X, 1914.

——: "On the partial correlation-ratio"; *Biom.*, Vol. XI, 1915.

——: "On certain probable errors and correlation coefficients of multiple frequency distributions with skew regression"; *Biom.*, Vol. XI, 1916.

——: "On the representation of statistical data"; *Biom.*, Vol. XI, 1917.

——: "On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables"; *Biom.*, Vol. XII, 1918.

——: "Formulae for determining the mean values of products of deviations of mixed moment coefficients in two to eight variables in samples taken from a limited population"; *Biom.*, Vol. XII, 1918.

KAPTEYN, J. C.: *Skew frequency-curves in biology and statistics; Gröningen*, 1903.

KELLEY, TRUMAN L.: *Educational Guidance;* Teachers College, Columbia University, 1914.

——: "Comparable measures"; *Jour. Ed. Psyc.*, 1914.

——: "A simplified method of using scaled data for purposes of testing"; *Sch. and Soc.*, Vol. IV, Nos. 79 and 80; 1916.

——: Tables to facilitate the calculation of partial coefficients of correlation and regression equations; *Bul. of the Univ. of Texas* (out of print), No. 27, May, 1916.

——: "Individual testing with completion test exercises"; *Teachers College Record*, Vol. XVIII, No. 4, Sept. 1917.

——: "Measurement of overlapping"; *Jour. of Educ. Psyc.*, Vol. X, No. 9, 1919.

——: "Principles underlying the classification of men": *Jour. of Applied Psyc.*, Vol. III, March, 1919.

——: *Chart to facilitate the calculation of partial coefficients of correlation and regression equations* (containing one 7 x 10 chart); Stanford University, 1921.

——: *Alignment chart of correlation functions, 17 × 23*, a supplement to the preceding, 1921.

KELLEY, TRUMAN L.: "Certain properties of index numbers"; *Quart., Am. Stat. Assn.*, Vol. XVII, Sept., 1921.

——: "The reliability of test scores"; *Jour. Ed. Res.*, May, 1921.

——: "A new measure of dispersion"; *Quart. Am. Stat. Assn.*, June, 1921.

KELLEY, TRUMAN L., and TERMAN, LEWIS M.: "Dr. Ruml's criticism of mental test methods"; *Jour. Philosophy*, Vol. XVIII, Aug., 1921.

KEYNES, J. N.: *A Treatise on Probability;* Macmillan, 1921.

KING, WILFORD ISBELL: *The Elements of Statistical Methods;* Macmillan, 1912.

KNIBBS, G. H.: "Prices, price indexes and cost of living in Australia," Bur. of Census and Stat., *Labor and Indust. Br. Report No. 1*, 1912. Also report No. 9, 1918.

KOREN, JOHN: *The History of Statistics;* Macmillan, 1918.

LAPLACE, PIERRE SIMON, MARQUIS DE: *Essai philosophique sur les probabilités*, 1814.

——: *Théorie analytique des probabilités*, 2d edition. 1814.

LEE, ALICE: "Tables of F (r, ν) and H(r, ν) Functions"; *British Assn. Report*, 1899.

——: "Table of the Gaussian 'tail' functions; when the 'tail' is larger than the body"; *Biom.*, Vol. X, 1914.

——: "Further supplementary tables for determining high correlations from tetrachoric groupings"; *Biom.*, Vol. XI, 1917.

LEE, ALICE: See Pearson and Lee (1908).

LEXIS, WILHELM: *Zur Theorie der Massenscheinunger in der menschlichen Gesellschaft.* Freiburg, 1877.

——: "Ueber die Theorie der Stabilität Statischen Reihen," *Jahrbücher f. Nationalökonomie u. Statistik*, Vol. XXXII, 1879.

——: *Abhandlungen zur Theorie der Bevölkerungs- und Moralstatistik;* Fischer, Jena, 1903.

MCEWEN, GEORGE F. and MICHAEL, ELLIS L.: "The functional relation of one variable to each of a number of correlated variables determined by a method of successive approximations to group averages: A contribution to statistical methods"; *Proc. of the Am. Acad. of Arts and Sciences*, Vol. LV, No. 2, Dec., 1919.

MARCH, L.: "Comparaison numérique de courbes statistiques"; *Jour. de la Soc. de Stat. de Paris*, Vol. XLVI, 1905.

MINER, JAMES BURT: "Correlation"; *Psyc. Bul.*, Vol. XIII, No. 5, 1916; Vol. XIV, No. 5, 1917; Vol. XV, No. 4, 1918; Vol. XVI, No. 11, 1919; Vol. XVII, No. 11, 1920.

MINER, JOHN RICE: Tables of $\sqrt{1 - r^2}$ and $1 - r^2$ for use in partial correlation and in trigonometry. Johns Hopkins Press, Baltimore, 1922.

MITCHELL, WESLEY C.: *Business cycles; California University Memoires*, Vol. III, 1913.

——: Author of part 1, *The making and use of index numbers*, of the Bulletin of the U. S. Bureau of Labor Statistics, whole number 173; 1915.

MOORE, CHARLES N.: "On the coefficient of correlation as a measure of relationship"; *Science*, Vol. XLII, No. 1086; 1915.

OTIS, ARTHUR S.: "The Reliability of spelling scales, including a 'deviation formula' for correlation"; *Sch. and Soc.*, Vol. IV, Nos. 96–99, 1916.

——: "A criticism of the Yerkes-Bridges point scale, with alternative suggestions"; *Jour. Ed. Psyc.*, Vol. VIII, No. 3, March, 1917.

——: "An absolute point scale for the group measurement of intelligence"; *Jour. Ed. Psy.*, Vol. IX, Nos. 5 and 6, May and July, 1918.

PAIRMAN, ELEANOR and PEARSON, KARL: "On corrections for the moment-coefficients of limited range frequency distributions when there are finite or infinite ordinates and any slopes at the terminals of the range"; *Biom.*, Vol. XII, 1919.

PEARL, R.: "Frequency constants of a variable $z = f(x_1, x_2)$"; *Biom.*, Vol. VI, 1909.

PEARSON, KARL: "Contributions to the mathematical theory of evolution (on the dissection of asymmetrical frequency-curves)"; *Phil. Trans. Roy. Soc.*, Series A, Vol. CLXXXV, p. 71; 1894.

——: "Skew variation in homogeneous material"; *Phil. Trans. A.*, Vol. CLXXXVI, pp. 343, et seq., 1895; and a supplement in *Phil. Trans. A.*, Vol. CXCVII, pp. 443–459, 1901.

——: "Regression, hereditary and panmixia"; *Phil. Trans. A.*, Vol. CLXXXVII, pp. 253–318, 1896.

——: "On a form of spurious correlation which may arise when indices are used"; etc. *Proc. Roy. Soc.*, Vol. LX, pp. 489–498; 1897.

——: "On the correlation of characters not quantitatively measurable"; *Phil. Trans. A.*, Vol. CXCV, pp. 1–47; 1900.

——: "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have risen from random sampling"; *Phil. Mag.*, Vol. L, 1900.

——: "On the lines and planes of closest fit to systems of points in space"; *Phil. Mag.*, 1901.

——: "On the influence of natural selection on the variability and correlation of organs"; *Phil. Trans. Roy. Soc. of London, A.*, Vol. CC, pp. 1–66; 1902.

——: On the mathematical theory of errors of judgment, with special reference to the personal equation. *Phil. Trans. Roy. Soc. of London, A.*, Vol. CXCVIII, pp. 235–299; 1902.

——: "On the systematic fitting of curves to observations and measurements"; *Biom.*, Vols. I and II, 1902.

——: "On a general theory of the method of false position"; *Phil. Mag.*, 1903.

——: "On the probable errors of frequency constants"; *Biom.*, Vol. II, 1903.

——: "On the theory of contingency and its relation to association and normal correlation," *Math. Contrib. to the Theory of Evolution*, Biometric Laboratory Publications, University of London; Cambridge University Press, 1904.

PEARSON, KARL: "On an elementary proof of Sheppard's formulæ for correcting raw moments and on other allied points"; *Biom.*, Vol. III, 1904.

——: "On the theory of skew correlation and non-linear regression," *Math. Contrib. to the Theory of Evolution*, Biometric Laboratory Publications; University of London, Cambridge University Press, 1905.

——: "On the curves which are most suitable for describing the frequency of random samples of a population"; *Biom.*, Vol. V, 1906.

——: "On certain points connected with scale order in the case of correlation of two characters, which for some arrangement give a linear regression line"; *Biom.*, Vol. V, 1906.

——: "On the significant or non-significant character of a sub-sample drawn from a sample"; *Biom.*, Vol. V, 1906.

——: "Skew frequency curves"; *Biom.*, Vol. V, 1906.

——: "On further methods of measuring correlation. Math. Contrib. to the Theory of Evolution"; *Biometric Laboratory Publications*, University of London; Cambridge University Press, 1907.

——: "Reply to certain criticisms of Mr. G. U. Yule"; *Biom.*, Vol. V, 1907.

——: "On the influence of double selection on the variation and correlation of two characters"; *Biom.*, Vol. VI, 1908.

——: "On a formula for determining $\Gamma(x+1)$"; *Biom.*, Vol. VI, 1908.

——: "On a new method of determining correlation between a measured character A and a character B, of which only the percentage of cases wherein B exceeds or falls short of a given intensity is recorded for each grade of A"; *Biom.*, Vol. VII, 1909.

——: "On the constants of index-distributions as deduced from the like constants for the components of the ratio, with special reference to the opsonic index"; *Biom.*, Vol. VII, 1910.

——: "On a new method of determining correlation, when one variable is given by alternative and the other by multiple categories"; *Biom.*, Vol. VII, 1910.

——: "On a correction needful in the case of the correlation ratio"; *Biom.*, Vol. VIII, 1911.

——: "Further remarks on the law of ancestral heredity"; *Biom.*, Vol. VIII, 1911.

——: The opsonic index. "Mathematical error and functional error." *Biom.*, Vol. VIII, 1911.

——: "On the probability that two independent distributions of frequency are really samples from the same population"; *Biom.*, Vol. VIII, 1911.

——: "On the general theory of the influence of selection on correlation and variation"; *Biom.*, Vol. VIII, 1912.

——: "On a novel method of regarding the association of two variates classed solely in alternative categories." *Math. Contrib. to the Theory of Evolution; Biometric Laboratory Publications*, University of London, Cambridge University Press, 1912.

PEARSON, KARL: "On the measurement of the influence of 'Broad Categories' on correlation"; *Biom.*, Vol. IX, 1913.

——: "On the probable errors of frequency constants"; *Biom.*, Vol. IX, 1913.

——: "On the probable error of a coefficient of correlation as found from a fourfold table"; *Biom.*, Vol. IX, 1913.

——: "On the surface of constant association $Q = 0.6$"; *Biom.*, Vol. IX, 1913.

——: "On certain errors with regard to multiple correlation occasionally made by those who have not adequately studied the subject"; *Biom.*, Vol. X, 1914.

——: "On an extension of the method of correlation by grades or ranks"; *Biom.*, Vol. X, 1914.

——: "On the probability that two independent distributions of frequency are really samples of the same population, with special reference to recent work on the identity of trypanosome strains"; *Biom.*, Vol. X, 1914.

——: "On certain types of compound frequency distributions in which the components can be individually described by binomial series"; *Biom.*, Vol. XI, 1915.

——: "On the probable error of a coefficient of mean square contingency"; *Biom.*, Vol. X, 1915.

——: "On the application of 'Goodness of Fit' tables to test regression curves and theoretical curves used to describe observational or experimental data"; *Biom.*, Vol. XI, 1916.

——: "On the general theory of multiple contingency with special reference to partial contingency"; *Biom.*, Vol. XI, 1916.

——: "On some novel properties of partial and multiple correlation coefficients in a universe of manifold characteristics"; *Biom.*, Vol. XI, 1916.

——: "On the probable error of biserial η"; *Biom.*, Vol. XI, 1917.

——: "The probable error of a Mendelian class frequency"; *Biom.*, Vol. XI, 1917.

——: "On generalised Tchebycheff Theorems in the mathematical theory of statistics"; *Biom.*, Vol. XII, 1919.

——: "Notes on the history of correlation"; *Biom.*, Vol. XIII, 1920.

——: "The fundamental problem of practical statistics"; *Biom.*, Vol. XIII, 1920.

——: "On the probable errors of frequency constants"; *Biom.*, Vol. XIII, 1920.

——: "On a general method of determining successive terms in a skew regression line"; *Biom.*, Vol. XIII, 1921.

——: Note on the "Fundamental problem of practical statistics"; *Biom.*, Vol. XIII, 1921.

PEARSON, KARL (Editor): *Tables for Statisticians and Biometricians*, Cambridge University Press, London, 1914.

PEARSON, KARL and BLAKEMAN, JOHN: "On the mathematical theory of random migration," *Math. Contrib. to the Theory of Evolution,* Biometric Laboratory Publications, University of London; Cambridge University Press, 1906.

PEARSON, KARL: See Filon and Pearson. 1898.

——: See also Blakeman and Pearson, 1906.

——: See also Cave and Pearson, 1914.

——: See Pairman and Pearson, 1919.

PEARSON, KARL and HERON, DAVID: "On theories of association"; *Biom.*, Vol. IX, 1913.

PEARSON, KARL and LEE, ALICE: "On the generalized probable error in multiple normal correlation"; *Biom.*, Vol. VI, 1908.

PEARSON and YOUNG, A. W.: "On the product-moments of various orders of the normal correlation surface of two variates"; *Biom.*, Vol. XII, 1918.

PEIRCE, B. O.: *A short table of integrals;* 2d rev. ed., Ginn, 1910.

PERRIN, EMILY: "On the contingency between occupation in the case of fathers and sons"; *Biom.*, Vol. III, 1904.

PERSONS, WARREN M.: "Construction of a business barometer based upon annual data"; *Am. Econ. Rev.*, Vol. 6, 1916.

——: "On the variate difference-correlation method and curve fitting"; *Quart. Am. Statis. Assn.*, Vol. 16, No. 118, 1917.

PETERS, J.: *Neue Rechentafeln für multiplikation und division* (gives products up to 100 × 10000); G. Reimer, Berlin.

PINTNER, RUDOLPH: "Comparison of the Ayres and Thorndike handwriting scales"; *Jour. Ed. Psych.*, Vol. V, No. 9, 1914.

POINCARÉ, H.: *Calcul des probabilités;* Gauthier-Villars, Paris, 1896.

POISSON, S. D.: "Sur la proportion des naissances des filles et des garçons"; *Mémoires de l'Acad. des Sciences,* Vol. IX, p. 239; 1829.

——: *Recherches sur la probabilité des jugements, etc.* Paris, 1837.

QUETELET, L. A. J.: *Lettres sur la théorie des probabilités, appliquée aux sciences morales et politiques;* 1846. (English translation by O. G. Downes, 1849.)

RHIND, A.: "Tables to facilitate the computation of the probable errors of the chief constants of skew frequency distributions"; *Biom.*, Vol. VII, 1909–10.

RIETZ, H. L.: "On functional relations for which the coefficient of correlation is zero"; *Quar. Am. Stat. Assn.*, 1919.

——: "Urn schemata as a basis for the development of correlation theory"; *Annals of Math.*, Vol. XXI, 1920.

RITCHIE-SCOTT, A.: "Note on the probable error of the coefficient of correlation in the variate difference correlation method"; *Biom.*, Vol. XI, 1915.

——: "The correlation coefficient of a polychoric table"; *Biom.*, Vol. XII, 1918.

RUGG, HAROLD ORDWAY: *Statistical Methods Applied to Education;* Houghton, Mifflin, 1917.

RUML, BEARDSLEY: "Measure of the efficiency of mental tests"; *Psy. Rev.*, Vol. XXIII, No. 6, 1916.

SECRIST, HORACE: *An Introduction to Statistical Methods;* Macmillan, 1917.

SHEPPARD, W. F.: "On the application of the theory of error to cases of normal distribution and normal correlation"; *Phil. Trans. A.*, Vol. CXCII, pp. 101–167, 1898.

——: "On the calculation of the most probable values of frequency constants for data arranged according to equi-distant divisions of a scale"; *Proc. Lon. Math. Soc.*, Vol. XXIX, pp. 353–380; 1898.

——: "On the calculation of the double-integral expressing normal correlation"; *Cambridge Phil. Trans.*, Vol. XIX, p. 23, 1900.

——: "New tables of the probability integral"; *Biom.*, Vol. II, 1903.

——: "The calculation of the moments of a frequency-distribution"; *Biom.*, Vol. V, 1907.

SMITH, KIRSTINE: "On the 'best' values of the constants in frequency distributions"; *Biom.*, Vol. XI, 1916.

——: "On the standard deviations of adjusted and interpolated values of an observed *Polynomial Function* and its constants and the guidance they give towards a proper choice of the distribution of observations"; *Biom.*, Vol. XI, 1917.

SNOW, E. C.: "Application of the method of multiple correlation to the establishment of post-censual populations"; *Jour. Roy. Stat. Soc.*, Vol. LXXIV, pp. 575–629, London, 1911.

——: "Application of the correlation coefficient to Mendelian distributions"; *Biom.*, Vol. VIII, 1912.

SOMMERVILLE, D. M. Y.: "On the classification of frequency ratios": *Biom.*, Vol. V, 1906.

SOPER, H. E.: "On the probable error of the correlation coefficient to a second approximation"; *Biom.*, Vol. IX, 1913.

——: "On the probable error of the bi-serial expression for the correlation coefficient"; *Biom.*, Vol. X, 1914.

——: "Tables of Poisson's exponential binomial limit": *Biom.*, Vol. X, 1914.

SOPER, H. E., YOUNG, A. W., CAVE, B. M., LEE, ALICE and PEARSON, KARL: "On the distribution of the correlation coefficient in small samples": *Biom.*, Vol. XI, 1917.

SPEARMAN, C.: "The proof and measurement of association between two things"; *Amer. Jour. of Psyc.*, Vol. XV, 1904.

——: "A footrule for measuring correlation"; *Brit. Jour. of Psyc.*, Vol. II, p. 89; 1906.

——: "Demonstration of formulæ for true measurement of correlation"; *Am. Jour. Psyc.*, Vol. XVIII, 1907.

——: "Coefficient of correlation calculated from faulty data"; *Brit. Jour. of Psy.*, Vol. III, 1910.

——: "Correlations of sums and differences"; *Brit. Jour. of Psy.*, Vol. V, 1913.

STUDENT: "Probable error of a correlation coefficient"; *Biom.*, Vol. VI, 1908.

STUDENT: "Probable error of a mean"; *Biom.*, Vol. VI, 1908.

——: "On the distribution of the means of samples which are not drawn at random"; *Biom.*, Vol. 7, 1909.

——: "The correction to be made to the correlation ratio for grouping"; *Biom.*, Vol. IX, 1913.

——: "The elimination of spurious correlation due to position in time or space"; *Biom.*, Vol. X, 1914.

——: "Tables for estimating the probability that the mean of a unique sample of observations lies between $-\infty$ and any given distance of the mean of the population from which the sample is drawn"; *Biom.*, Vol. XI, 1917.

——: "An explanation of deviations from Poisson's law in practice"; *Biom.*, Vol. XII, 1919.

——: "An experimental determination of the probable error of Dr. Spearman's correlation coefficients"; *Biom.*, Vol. XIII, 1921.

TCHOUPROFF, AL. A.: "On the mathematical expectation of the moments of frequency distributions"; *Biom.*, Vol. XII, Part One, 1918; Part Two, 1919; *Biom.*, Vol. XIII; Part Three, 1921.

TERMAN, LEWIS M.: "The intelligence quotient of Francis Galton in childhood"; *Am. Jour. of Psy.*, Vol. XXVIII, 1917.

TERMAN, LEWIS M.: See also Kelley and Terman, 1921.

THIELE, T. N.: *Theory of Observations;* London, 1903.

THOMSON, GODFREY H.: "The criterion of goodness of fit of psychophysical curves"; *Biom.*, Vol. XII, 1919.

——: "A direct deduction of the constant process used in the method of right and wrong cases"; *Psych. Rev.*, Vol. XXVI, No. 6, 1919.

——: "On the degree of perfection of hierarchical order among correlation coefficients"; *Biom.*, Vol. XII, 1919.

——: See also Brown and Thompson, 1921.

THORNDIKE, EDWARD L.: *Empirical Studies in the Theory of Measurement;* Archives of Psyc. (New York), 1907.

——: *Mental and Social measurements;* Teachers College, Columbia University. 1913.

THURSTONE, L. L.: "A scoring method for mental tests"; *Psy. Bul.*, Vol. XVI, No. 7, 1919.

TODHUNTER, I. A.: *History of the Mathematical Theory of Probability from the Time of Pascal to that of Laplace;* Macmillan, 1865.

URBAN, F. M.: "The application of statistical methods to the problems of psychophysics"; *Exper. Studies in Psyc. and Ped.*, III, Philadelphia, 1908.

——: "Die psychophysischen Massmethoden als Grundlagen empirischer Messungen"; *Archiv f. d. ges. Psychol.*, Vol. XVI, 1909.

——: *Die Praxis der Konstanzmethode;* Leipzig. 1912.

VENN, J.: *The Logic of Chance:* an essay on the foundations and province of the theory of probability, with especial reference to its logical bearings and its application to moral and social science and to statistics. Macmillan & Co., London, Third ed.; 1888.

WALSH, C. M.: *Measurement of General Exchange Value.* Macmillan, 1901.

——: *Problem of Estimation;* P. S. King, London, 1921.

WEST, C. J.: *Introduction to Mathematical Statistics;* R. G. Adams & Co., Columbus, Ohio, 1918.

WESTERGAARD, H.: *Die Grundzüge der Theorie der Statistik;* Fischer, Jena, 1890.

WHIPPLE, GEORGE CHANDLER. *Vital statistics;* Wiley, 1919.

WHIPPLE, G. M.: *Manual of Mental and Physical Tests;* Sec. Ed., Two Parts; Warwick and York. 1914.

WHITAKER, LUCY: *"On Poisson's law of small numbers"; Biom.*, Vol. X; 1914.

WOODWORTH, ROBERT SESSIONS: "Combining the results of several tests: a study in statistical method"; *Psyc. Rev.*, Vol. XIX, 1912.

WRIGHT, THOMAS W. and HAYFORD, JOHN F.: *Adjustment of Observations by the Method of Least Squares with Applications to Geodetic Work;* D. Van Nostrand Co., 1906.

YERKES, ROBERT M. (Editor): "Psychological examining in the United States army"; *Nat'l Acad. of Science*, Vol. XV, 1921.

YOUNG, ANDREW W.: "Note on the standard deviations of samples of two or three"; *Biom.*, Vol. XI, 1916.

YOUNG, ANDREW W. and PEARSON, KARL: "On the probable error of a coefficient of contingency without approximation"; *Biom.*, Vol. XI, 1916.

YULE, G. U.: "Notes on the history of pauperism"; *Jour. of Roy. Stat. Soc.*, Vol. LIX, pp. 318–357; 1896.

——: "On the significance of Bravais' formulæ for regression, etc., in the case of skew correlation"; *Proc. Roy. Soc.*, Vol. LX, p. 477, 1897.

——: "On the theory of correlation"; *Jour. Roy. Stat. Soc.*, Vol. LX, 1897.

——: "On the association of attributes in statistics"; *Phil. Trans. Roy. Soc.*, Series A, Vol. CXCIV, p. 257, 1900.

——: Notes on the theory of association of attributes in statistics. *Biom.* Vol. 2, 1903.

——: "On the theory of correlation for any number of variables treated by a new system of notation"; *Proc. Roy. Soc.*, Series A, Vol. LXXIX, p. 182, 1907.

——: "On the interpretation of correlations between indices or ratios"; *Jour. Roy. Stat. Soc.*, Vol. LXXIII, p. 644, 1910.

——: "On the methods of measuring the association between two attributes"; *Jour. Roy. Stat. Soc.*, Vol. LXXV, 1912.

——: *Introduction to the Theory of Statistics;* Lippincott, 1912.

YULE, G. U.: See Hooker, R. H. and Yule, G. U., 1906.

ZIMMERMAN, H.: *Rechentafel, nebst Sammlung häufig gebrauchter Zahlenwerthe;* W. Ernst and Son, Berlin. Eng. edition, Asher and Co., London.

APPENDIX C

KELLEY–WOOD TABLE OF THE NORMAL PROBABILITY INTEGRAL

.000			.500				.500
I	x	z	q	z/q	z/p	pq	p
.000	.000000	.398942	.500	.79788	.79788	.250000	.500
.001	.002507	.398941	.499	.79948	.79629	.249999	.501
.002	.005013	.398937	.498	.80108	.79470	.249996	.502
.003	.007520	.398931	.497	.80268	.79310	.249991	.503
.004	.010027	.398922	.496	.80428	.79151	.249984	.504
.005	.012533	.398911	.495	.80588	.78992	.249975	.505
.006	.015040	.398897	.494	.80748	.78833	.249964	.506
.007	.017547	.398881	.493	.80909	.78675	.249951	.507
.008	.020054	.398862	.492	.81070	.78516	.249936	.508
.009	.022562	.398841	.491	.81230	.78358	.249919	.509
.010	.025069	.398816	.490	.81391	.78199	.249900	.510
.011	.027576	.398791	.489	.81552	.78041	.249879	.511
.012	.030084	.398762	.488	.81714	.77883	.249856	.512
.013	.032592	.398730	.487	.81875	.77725	.249831	.513
.014	.035100	.398697	.486	.82036	.77567	.249804	.514
.015	.037608	.398660	.485	.82198	.77410	.249775	.515
.016	.040117	.398621	.484	.82360	.77252	.249744	.516
.017	.042626	.398580	.483	.82522	.77095	.249711	.517
.018	.045135	.398536	.482	.82684	.76938	.249676	.518
.019	.047644	.398490	.481	.82846	.76780	.249639	.519
.020	.050154	.398441	.480	.83008	.76623	.249600	.520
.021	.052664	.398389	.479	.83171	.76466	.249559	.521
.022	.055174	.398336	.478	.83334	.76309	.249516	.522
.023	.057684	.398279	.477	.83497	.76153	.249471	.523
.024	.060195	.398220	.476	.83660	.75996	.249424	.524
.025	.062707	.398159	.475	.83823	.75840	.249375	.525
.026	.065219	.398096	.474	.83986	.75683	.249324	.526
.027	.067731	.398028	.473	.84150	.75527	.249271	.527
.028	.070243	.397959	.472	.84292	.75371	.249216	.528
.029	.072756	.397888	.471	.84477	.75215	.249159	.529
.030	.075270	.397814	.470	.84641	.75059	.249100	.530
.031	.077784	.397737	.469	.84805	.74903	.249039	.531
.032	.080298	.397658	.468	.84970	.74748	.248976	.532
.033	.082813	.397577	.467	.85134	.74592	.248911	.533
.034	.085329	.397493	.466	.85299	.74437	.248844	.534
.035	.087845	.397406	.465	.85464	.74281	.248775	.535
.036	.090361	.397317	.464	.85629	.74126	.248704	.536
.037	.092878	.397225	.463	.85794	.73971	.248631	.537
.038	.095395	.397131	.462	.85959	.73816	.248556	.538
.039	.097914	.397034	.461	.86125	.73661	.248479	.539
.040	.100434	.396935	.460	.86290	.73506	.248400	.540
.040			.460				.540

.040 .460 .540

I	x	z	q	z/q	z/p	pq	p
.040	.100434	.396935	.460	.86290	.73506	.248400	.540
.041	.102953	.396834	.459	.86456	.73352	.248319	.541
.042	.105474	.396729	.458	.86622	.73197	.248236	.542
.043	.107995	.396623	.457	.86788	.73043	.248151	.543
.044	.110516	.396513	.456	.86955	.72888	.248064	.544
.045	.113039	.396401	.455	.87121	.72734	.247975	.545
.046	.115562	.396287	.454	.87288	.72580	.247884	.546
.047	.118085	.396170	.453	.87455	.72426	.247791	.547
.048	.120610	.396051	.452	.87622	.72272	.247696	.548
.049	.123135	.395929	.451	.87789	.72118	.247599	.549
.050	.125661	.395805	.450	.87957	.71964	.247500	.550
.051	.128188	.395678	.449	.88124	.71811	.247399	.551
.052	.130716	.395549	.448	.88292	.71657	.247296	.552
.053	.133245	.395417	.447	.88460	.71504	.247191	.553
.054	.135774	.395282	.446	.88628	.71350	.247084	.554
.055	.138304	.395145	.445	.88797	.71197	.246975	.555
.056	.140835	.395005	.444	.88965	.71044	.246864	.556
.057	.143367	.394863	.443	.89134	.70891	.246751	.557
.058	.145900	.394719	.442	.89303	.70738	.246636	.558
.059	.148434	.394572	.441	.89472	.70585	.246519	.559
.060	.150969	.394422	.440	.89641	.70432	.246400	.560
.061	.153505	.394270	.439	.89811	.70280	.246279	.561
.062	.156042	.394115	.438	.89981	.70127	.246156	.562
.063	.158580	.393957	.437	.90150	.69975	.246031	.563
.064	.161119	.393798	.436	.90321	.69822	.245904	.564
.065	.163658	.393635	.435	.90491	.69670	.245775	.565
.066	.166199	.393470	.434	.90661	.69518	.245644	.566
.067	.168741	.393303	.433	.90832	.69366	.245511	.567
.068	.171285	.393133	.432	.91003	.69214	.245376	.568
.069	.173829	.392960	.431	.91174	.69062	.245239	.569
.070	.176374	.392785	.430	.91345	.68910	.245100	.570
.071	.178921	.392608	.429	.91517	.68758	.244959	.571
.072	.181468	.392427	.428	.91689	.68606	.244816	.572
.073	.184017	.392245	.427	.91861	.68455	.244671	.573
.074	.186567	.392059	.426	.92033	.68303	.244524	.574
.075	.189118	.391870	.425	.92205	.68151	.244375	.575
.076	.191671	.391681	.424	.92378	.68000	.244224	.576
.077	.194225	.391488	.423	.92550	.67849	.244071	.577
.078	.196780	.391293	.422	.92723	.67698	.243916	.578
.079	.199336	.391095	.421	.92897	.67547	.243759	.579
.080	.201893	.390894	.420	.93070	.67396	.243600	.580

.080 .420 .580

.080 .420 .580

I	x	z	q	z/q	z/p	pq	p
.080	.201893	.390894	.420	.93070	.67396	.243600	.580
.081	.204452	.390691	.419	.93244	.67245	.243439	.581
.082	.207013	.390485	.418	.93417	.67094	.243276	.582
.083	.209574	.390277	.417	.93592	.66943	.243111	.583
.084	.212137	.390066	.416	.93766	.66792	.242944	.584
.085	.214702	.389852	.415	.93940	.66641	.242775	.585
.086	.217267	.389636	.414	.94115	.66491	.242604	.586
.087	.219835	.389418	.413	.94290	.66340	.242431	.587
.088	.222403	.389197	.412	.94465	.66190	.242256	.588
.089	.224973	.388973	.411	.94641	66040	.242079	.589
.090	.227545	.388747	.410	.94816	.65889	.241900	.590
.091	.230118	.388518	.409	.94992	.65739	.241719	.591
.092	.232693	.388287	.408	.95168	.65589	.241536	.592
.093	.235269	.388053	.407	.95345	.65439	.241351	.593
.094	.237847	.387816	.406	.95521	.65289	.241164	.594
.095	.240426	.387577	.405	.95698	.65139	.240975	.595
.096	.243007	.387335	.404	.95875	.64989	.240784	.596
.097	.245590	.387091	.403	96052	.64839	.240591	.597
.098	.248174	.386844	.402	.96230	.64690	.240396	.598
.099	.250760	.386595	.401	.96408	.64540	.240199	.599
.100	.253347	.386342	.400	.96586	.64390	.240000	.600
.101	.255936	.386088	.399	.96764	.64241	.239799	.601
.102	.258527	.385831	.398	.96942	.64091	.239596	.602
.103	.261120	.385571	.397	.97121	.63942	.239391	.603
.104	.263714	.385308	.396	.97300	.63793	.239184	.604
.105	.266311	.385043	.395	.97479	.63749	.238975	.605
.106	.268909	.384776	.394	.97659	.63494	.238764	.606
.107	.271508	.384506	.393	.97839	.63345	.238551	.607
.108	.274110	.384233	.392	.98019	.63196	.238336	.608
.109	.276714	.383957	.391	.98199	.63047	.238119	.609
.110	.279319	.383679	.390	.98379	.62898	.237900	.610
.111	.281926	.383399	.389	.98560	.62749	.237679	.611
.112	.284536	.383115	.388	.98741	.62600	.237456	.612
.113	.287147	.382830	.387	.98922	.62452	.237231	.613
.114	.289760	.382541	.386	.99104	.62303	.237004	.614
.115	.292375	.382250	.385	.99286	.62154	.236775	.615
.116	.294992	.381956	.384	.99468	.62006	.236544	.616
.117	.297611	.381660	.383	.99650	.61857	.236311	.617
.118	.300232	.381361	.382	.99833	.61709	.236076	.618
.119	.302855	.381060	.381	1.00016	.61561	.235839	.619
.120	.305481	.380755	.380	1.00199	.61412	.235600	.620

.120 .380 .620

.120 .380 .620

I	x	z	q	z/q	z/p	pq	p
.120	.305481	.380755	.380	1.00199	.61412	.235600	.620
.121	.308108	.380449	.379	1.00382	.61264	.235359	.621
.122	.310738	.380139	.378	1.00566	.61116	.235116	.622
.123	.313369	.379827	.377	1.00750	.60967	.234871	.623
.124	.316003	.379513	.376	1.00934	.60819	.234624	.624
.125	.318639	.379195	.375	1.01119	.60671	.234375	.625
.126	.321278	.378875	.374	1.01303	.60523	.234124	.626
.127	.323918	.378553	.373	1.01489	.60375	.233871	.627
.128	.326561	.378227	.372	1.01674	.60227	.233616	.628
.129	.329206	.377900	.371	1.01860	.60079	.233359	.629
.130	.331853	.377569	.370	1.02046	.59932	.233100	.630
.131	.334503	.377236	.369	1.02232	.59784	.232839	.631
.132	.337155	.376900	.368	1.02418	.59636	.232576	.632
.133	.339809	.376562	.367	1.02605	.59488	.232311	.633
.134	.342466	.376220	.366	1.02792	.59341	.232044	.634
.135	.345125	.375877	.365	1.02980	.59193	.231775	.635
.136	.347787	.375530	.364	1.03168	.59046	.231504	.636
.137	.350451	.375181	.363	1.03356	.58898	.231231	.637
.138	.353118	.374829	.362	1.03544	.58751	.230956	.638
.139	.355787	.374475	.361	1.03733	.58603	.230679	.639
.140	.358459	.374118	.360	1.03922	.58456	.230400	.640
.141	.361133	.373758	.359	1.04111	.58309	.230119	.641
.142	.363810	.373395	.358	1.04300	.58161	.229836	.642
.143	.366489	.373030	.357	1.04490	.58014	.229551	.643
.144	.369171	.372662	.356	1.04680	.57867	.229264	.644
.145	.371856	.372292	.355	1.04871	.57720	.228975	.645
.146	.374544	.371919	.354	1.05062	.57573	.228684	.646
.147	.377234	.371543	.353	1.05253	.57426	.228391	.647
.148	.379927	.371164	.352	1.05444	.57278	.228096	.648
.149	.382622	.370783	.351	1.05636	.57131	.227799	.649
.150	.385320	.370399	.350	1.05828	.56984	.227500	.650
.151	.388022	.370012	.349	1.06021	.56837	.227199	.651
.152	.390726	.369623	.348	1.06214	.56691	.226896	.652
.153	.393433	.369231	.347	1.06407	.56544	.226591	.653
.154	.396142	.368836	.346	1.06600	.56397	.226284	.654
.155	.398855	.368439	.345	1.06794	.56250	.225975	.655
.156	.401571	.368038	.344	1.06988	.56103	.225664	.656
.157	.404289	.367635	.343	1.07182	.55957	.225351	.657
.158	.407011	.367230	.342	1.07377	.55810	.225036	.658
.159	.409735	.366821	.341	1.07572	.55663	.224719	.659
.160	.412463	.366410	.340	1.07768	.55517	.224400	.660

.160 .340 .660

.160 .340 .660

I	x	z	q	z/q	z/p	pq	p
.160	.412463	.366410	.340	1.07768	.55517	.224400	.660
.161	.415194	.365996	.339	1.07963	.55370	.224079	.661
.162	.417928	.365580	.338	1.08160	.55224	.223756	.662
.163	.420665	.365160	.337	1.08356	.55077	.223431	.663
164	.423405	.364738	.336	1.08553	.54930	.223104	.664
.165	.426148	.364314	.335	1.08750	.54784	.222775	.665
.166	.428895	.363886	.334	1.08948	.54638	.222444	.666
.167	.431644	.363456	.333	1.09146	.54491	.222111	.667
.168	.434397	.363023	.332	1.09344	.54345	.221776	.668
.169	.437154	.362587	.331	1.09543	.54198	.221439	.669
.170	.439913	.362149	.330	1.09742	.54052	.221100	.670
.171	.442676	.361707	.329	1.09941	.53906	.220759	.671
.172	.445443	.361263	.328	1.10141	.53759	.220416	.672
.173	.448212	.360817	.327	1.10342	.53613	.220071	.673
.174	.450986	.360367	.326	1.10542	.53467	.219724	.674
.175	.453762	.359915	.325	1.10743	.53321	.219375	.675
.176	.456542	.359459	.324	1.10944	.53174	.219024	.676
.177	.459326	.359001	.323	1.11146	.53028	.218671	.677
.178	.462113	.358541	.322	1.11348	.52882	.218316	.678
.179	.464904	.358077	.321	1.11550	.52736	.217959	.679
.180	.467699	.357611	.320	1.11753	.52590	.217600	.680
.181	.470497	.357142	.319	1.11957	.52444	.217239	.681
.182	.473299	.356670	.318	1.12160	.52298	.216876	.682
.183	.476104	.356195	.317	1.12364	.52152	.216511	.683
.184	.478914	.355718	.316	1.12569	.52006	.216144	.684
.185	.481727	.355237	.315	1.12774	.51859	.215775	.685
.186	.484544	.354754	.314	1.12979	.51713	.215404	.686
.187	.487365	.354268	.313	1.13185	.51567	.215031	.687
.188	.490189	.353780	.312	1.13391	.51422	.214656	.688
.189	.493018	.353288	.311	1.13597	.51275	.214279	.689
.190	.495850	.352793	.310	1.13804	.51129	.213900	.690
.191	.498687	.352296	.309	1.14012	.50984	.213519	.691
.192	.501527	.351796	.308	1.14219	.50838	.213136	.692
.193	.504372	.351293	.307	1.14428	.50692	.212751	.693
.194	.507221	.350787	.306	1.14636	.50546	.212364	.694
.195	.510073	.350279	.305	1.14846	.50400	.211975	.695
.196	.512930	.349767	.304	1.15055	.50254	.211584	.696
.197	.515792	.349253	.303	1.15265	.50108	.211191	.697
.198	.518657	.348736	.302	1.15475	.49962	.210796	.698
.199	.521527	.348216	.301	1.15686	.49816	.210399	.699
.200	.524401	.347693	.300	1.15898	.49670	.210000	.700

.200 .300 .700

.200 .300 .700

I	x	z	q	z/q	z/p	pq	p
.200	.524401	.347693	.300	1.15898	.49670	.210000	.700
.201	.527279	.347167	.299	1.16109	.49525	.209599	.701
.202	.530161	.346638	.298	1.16321	.49379	.209196	.702
.203	.533048	.346107	.297	1.16534	.49233	.208791	.703
.204	.535940	.345572	.296	1.16747	.49087	.208384	.704
.205	.538836	.345035	.295	1.16961	.48941	.207975	.705
.206	.541737	.344494	.294	1.17175	.48795	.207564	.706
.207	.544642	.343951	.293	1.17389	.48649	.207151	.707
.208	.547551	.343405	.292	1.17604	.48504	.206736	.708
.209	.550466	.342856	.291	1.17820	.48358	.206319	.709
.210	.553385	.342304	.290	1.18036	.48212	.205900	.710
.211	.556308	.341749	.289	1.18252	.48066	.205479	.711
.212	.559237	.341191	.288	1.18469	.47920	.205056	.712
.213	.562170	.340631	.287	1.18687	.47774	.204631	.713
.214	.565108	.340067	.286	1.18904	.47628	.204204	.714
.215	.568051	.339500	.285	1.19123	.47483	.203775	.715
.216	.570999	.338931	.284	1.19342	.47337	.203344	.716
.217	.573952	.338358	.283	1.19561	.47191	.202911	.717
.218	.576910	.337783	.282	1.19781	.47045	.202476	.718
.219	.579873	.337205	.281	1.20002	.46899	.202039	.719
.220	.582841	.336623	.280	1.20223	.46753	.201600	.720
.221	.585815	.336039	.279	1.20444	.46607	.201159	.721
.222	.588793	.335452	.278	1.20666	.46461	.200716	.722
.223	.591777	.334861	.277	1.20888	.46315	.200271	.723
.224	.594766	.334268	.276	1.21112	.46170	.199824	.724
.225	.597760	.333672	.275	1.21335	.46024	.199375	.725
.226	.600760	.333073	.274	1.21559	.45878	.198924	.726
.227	.603765	.332470	.273	1.21784	.45732	.198471	.727
.228	.606775	.331865	.272	1.22009	.45586	.198016	.728
.229	.609791	.331257	.271	1.22235	.45440	.197559	.729
.230	.612813	.330646	.270	1.22461	.45294	.197100	.730
.231	.615840	.330031	.269	1.22688	.45148	.196639	.731
.232	.618873	.329414	.268	1.22916	.45002	.196176	.732
.233	.621912	.328793	.267	1.23143	.44856	.195711	.733
.234	.624956	.328170	.266	1.23372	.44710	.195244	.734
.235	.628006	.327544	.265	1.23601	.44564	.194775	.735
.236	.631062	.326914	.264	1.23831	.44418	.194304	.736
.237	.634124	.326281	.263	1.24061	.44272	.193831	.737
.238	.637192	.325646	.262	1.24292	.44125	.193356	.738
.239	.640265	.325007	.261	1.24524	.43979	.192879	.739
.240	.643345	.324365	.260	1.24756	.43833	.192400	.740

.240 .260 .740

.240 .260 .740

I	x	z	q	z/q	z/p	pq	p
.240	.643345	.324365	.260	1.24756	.43833	.192400	.740
.241	.646431	.323720	.259	1.24988	.43687	.191919	.741
.242	.649524	.323072	.258	1.25222	.43541	.191436	.742
.243	.652622	.322421	.257	1.25456	.43394	.190951	.743
.244	.655727	.321767	.256	1.25690	.43248	.190464	.744
.245	.658838	.321110	.255	1.25925	.43102	.189975	.745
.246	.661955	.320449	.254	1.26161	.42956	.189484	.746
.247	.665079	.319786	.253	1.26398	.42809	.188991	.747
.248	.668209	.319119	.252	1.26635	.42663	.188496	.748
.249	.671346	.318449	.251	1.26872	.42517	.187999	.749
.250	.674490	.317777	.250	1.27111	.42370	.187500	.750
.251	.677640	.317101	.249	1.27350	.42224	.186999	.751
.252	.680797	.316421	.248	1.27589	.42077	.186496	.752
.253	.683961	.315739	.247	1.27830	.41931	.185991	.753
.254	.687131	.315053	.246	1.28070	.41784	.185484	.754
.255	.690309	.314365	.245	1.28312	.41638	.184975	.755
.256	.693493	.313673	.244	1.28554	.41491	.184464	.756
.257	.696685	.312978	.243	1.28798	.41345	.183951	.757
.258	.699884	.312279	.242	1.29041	.41198	.183436	.758
.259	.703090	.311578	.241	1.29285	.41051	.182919	.759
.260	.706303	.310873	.240	1.29531	.40904	.182400	.760
.261	.709523	.310165	.239	1.29776	.40758	.181879	.761
.262	.712751	.309454	.238	1.30023	.40611	.181356	.762
.263	.715986	.308740	.237	1.30270	.40464	.180831	.763
.264	.719229	.308022	.236	1.30518	.40317	.180304	.764
.265	.722479	.307301	.235	1.30767	.40170	.179775	.765
.266	.725737	.306577	.234	1.31016	.40023	.179244	.766
.267	.729003	.305850	.233	1.31266	.39876	.178711	.767
.268	.732276	.305119	.232	1.31517	.39729	.178176	.768
.269	.735558	.304385	.231	1.31768	.39582	.177639	.769
.270	.738847	.303648	.230	1.32021	.39435	.177100	.770
.271	.742144	.302908	.229	1.32274	.39288	.176559	.771
.272	.745450	.302164	.228	1.32528	.39140	.176016	.772
.273	.748763	.301417	.227	1.32783	.38993	.175471	.773
.274	.752085	.300666	.226	1.33038	.38846	.174924	.774
.275	.755415	.299913	.225	1.33294	.38698	.174375	.775
.276	.758754	.299155	.224	1.33551	.38551	.173824	.776
.277	.762101	.298395	.223	1.33809	.38403	.173271	.777
.278	.765456	.297631	.222	1.34068	.38256	.172716	.778
.279	.768820	.296864	.221	1.34328	.38108	.172159	.779
.280	.772193	.296094	.220	1.34588	.38069	.171600	.780

.280 .220 .780

.280			.220				.780
I	x	z	q	z/q	z/p	pq	p
.280	.772193	.296094	.220	1.34588	.38069	.171600	.780
.281	.775575	.295320	.219	1.34849	.37813	.171039	.781
.282	.778966	.294542	.218	1.35111	.37665	.170476	.782
.283	.782365	.293762	.217	1.35374	.37517	.169911	.783
.284	.785774	.292978	.216	1.35638	.37370	.169344	.784
.285	.789192	.292190	.215	1.35902	.37222	.168775	.785
.286	.792619	.291399	.214	1.36168	.37074	.168204	.786
.287	.796055	.290605	.213	1.36434	.36926	.167631	.787
.288	.799501	.289807	.212	1.36701	.36778	.167056	.788
.289	.802956	.289006	.211	1.36970	.36629	.166479	.789
.290	.806421	.288201	.210	1.37239	.36481	.165900	.790
.291	.809896	.287393	.209	1.37509	.36333	.165319	.791
.292	.813380	.286582	.208	1.37780	.36185	.164736	.792
.293	.816875	.285766	.207	1.38051	.36036	.164151	.793
.294	.820379	.284948	.206	1.38324	.35888	.163564	.794
.295	.823894	.284126	.205	1.38598	.35739	.162975	.795
.296	.827418	.283300	.204	1.38873	.35590	.162384	.796
.297	.830953	.282471	.203	1.39148	.35442	.161791	.797
.298	.834499	.281638	.202	1.39425	.35293	.161196	.798
.299	.838055	.280802	.201	1.39702	.35144	.160599	.799
.300	.841621	.279962	.200	1.39981	.34995	.160000	.800
.301	.845199	.279118	.199	1.40260	.34846	.159399	.801
.302	.848787	.278272	.198	1.40541	.34697	.158796	.802
.303	.852386	.277421	.197	1.40823	.34548	.158191	.803
.304	.855996	.276567	.196	1.41106	.34399	.157584	.804
.305	.859617	.275709	.195	1.41389	.34250	.156975	.805
.306	.863250	.274847	.194	1.41674	.34100	.156364	.806
.307	.866894	.273982	.193	1.41960	.33951	.155751	.807
.308	.870550	.273114	.192	1.42247	.33801	.155136	.808
.309	.874217	.272241	.191	1.42535	.33652	.154519	.809
.310	.877896	.271365	.190	1.42824	.33502	.153900	.810
.311	.881587	.270486	.189	1.43114	.33352	.153279	.811
.312	.885291	.269602	.188	1.43405	.33202	.152656	.812
.313	.889006	.268715	.187	1.43698	.33052	.152031	.813
.314	.892733	.267824	.186	1.43991	.32902	.151404	.814
.315	.896473	.266929	.185	1.44286	.32752	.150775	.815
.316	.900226	.266031	.184	1.44582	.32602	.150144	.816
.317	.903991	.265129	.183	1.44879	.32452	.149511	.817
.318	.907770	.264223	.182	1.45177	.32301	.148876	.818
.319	.911561	.263313	.181	1.45477	.32151	.148239	.819
.320	.915365	.262400	.180	1.45778	.32000	.147600	.820
.320			.180				.820

.320 .180 .820

I	x	z	q	z/q	z/p	pq	p
.320	.915365	.262400	.180	1.45778	.32000	.147600	.820
.321	.919183	.261483	.179	1.46080	.31849	.146959	.821
.322	.923014	.260562	.178	1.46383	.31699	.146316	.822
.323	.926859	.259637	.177	1.46688	.31548	.145671	.823
.324	.930717	.258708	.176	1.46993	.31397	.145024	.824
.325	.934589	.257775	.175	1.47300	.31245	.144375	.825
.326	.938476	.256839	.174	1.47609	.31094	.143724	.826
.327	.942376	.255898	.173	1.47918	.30943	.143071	.827
.328	.946291	.254954	.172	1.48229	.30792	.142416	.828
.329	.950221	.254006	.171	1.48541	.30640	.141759	.829
.330	.954165	.253054	.170	1.48855	.30488	.141100	.830
.331	.958125	.252097	.169	1.49170	.30337	.140439	.831
.332	.962099	.251137	.168	1.49486	.30185	.139776	.832
.333	.966088	.250173	.167	1.49804	.30033	.139111	.833
.334	.970093	.249205	.166	1.50123	.29881	.138444	.834
.335	.974114	.248233	.165	1.50444	.29728	.137775	.835
.336	.978150	.247257	.164	1.50766	.29576	.137104	.836
.337	.982203	.246277	.163	1.51090	.29424	.136431	.837
.338	.986271	.245292	.162	1.51415	.29271	.135756	.838
.339	.990356	.244304	.161	1.51742	.29118	.135079	.839
.340	.994458	.243312	.160	1.52070	.28966	.134400	.840
.341	.998576	.242315	.159	1.52399	.28813	.133719	.841
.342	1.002712	.241315	.158	1.52731	.28660	.133036	.842
.343	1.006864	.240310	.157	1.53064	.28507	.132351	.843
.344	1.011034	.239301	.156	1.53398	.28353	.131664	.844
.345	1.015222	.238288	.155	1.53734	.28200	.130975	.845
.346	1.019428	.237270	.154	1.54071	.28046	.130284	.846
.347	1.023651	.236249	.153	1.54411	.27892	.129591	.847
.348	1.027893	.235223	.152	1.54752	.27739	.128896	.848
.349	1.032154	.234193	.151	1.55095	.27585	.128199	.849
.350	1.036433	.233159	.150	1.55439	.27430	.127500	.850
.351	1.040732	.232120	.149	1.55785	.27276	.126799	.851
.352	1.045050	.231077	.148	1.56133	.27122	.126096	.852
.353	1.049387	.230030	.147	1.56483	.26967	.125391	.853
.354	1.053744	.228979	.146	1.56835	.26813	.124684	.854
.355	1.058122	.227923	.145	1.57188	.26658	.123975	.855
.356	1.062519	.226862	.144	1.57543	.26503	.123264	.856
.357	1.066938	.225798	.143	1.57901	.26347	.122551	.857
.358	1.071377	.224728	.142	1.58259	.26192	.121836	.858
.359	1.075837	.223655	.141	1.58621	.26037	.121119	.859
.360	1.080319	.222577	.140	1.58983	.25881	.120400	.860

.360 .140 .860

.360 .140 .860

I	x	z	q	z/q	z/p	pq	p
.360	1.080319	.222577	.140	1.58983	.25881	.120400	.860
.361	1.084823	.221494	.139	1.59348	.25725	.119679	.861
.362	1.089349	.220407	.138	1.59715	.25569	.118956	.862
.363	1.093897	.219315	.137	1.60084	.25413	.118231	.863
.364	1.098468	.218219	.136	1.60455	.25257	.117504	.864
.365	1.103063	.217119	.135	1.60829	.25100	.116775	.865
.366	1.107680	.216013	.134	1.61204	.24944	.116044	.866
.367	1.112321	.214903	.133	1.61581	.24787	.115311	.867
.368	1.116987	.213789	.132	1.61961	.24630	.114576	.868
.369	1.121676	.212669	.131	1.62343	.24473	.113839	.869
.370	1.126391	.211545	.130	1.62727	.24316	.113100	.870
.371	1.131131	.210416	.129	1.63113	.24158	.112359	.871
.372	1.135896	.209283	.128	1.63502	.24000	.111616	.872
.373	1.140688	.208145	.127	1.63894	.23842	.110871	.873
.374	1.145505	.207001	.126	1.64286	.23684	.110124	.874
.375	1.150349	.205853	.125	1.64683	.23526	.109375	.875
.376	1.155221	.204701	.124	1.65081	.23368	.108624	.876
.377	1.160120	.203543	.123	1.65482	.23209	.107871	.877
.378	1.165047	.202380	.122	1.65885	.23050	.107116	.878
.379	1.170002	.201213	.121	1.66292	.22891	.106359	.879
.380	1.174987	.200040	.120	1.66700	.22732	.105600	.880
.381	1.180001	.198863	.119	1.67112	.22572	.104839	.881
.382	1.185044	.197680	.118	1.67525	.22413	.104076	.882
.383	1.190118	.196493	.117	1.67943	.22253	.103311	.883
.384	1.195223	.195300	.116	1.68362	.22093	.102544	.884
.385	1.200359	.194102	.115	1.68785	.21932	.101775	.885
.386	1.205527	.192900	.114	1.69211	.21772	.101004	.886
.387	1.210727	.191691	.113	1.69638	.21611	.100231	.887
.388	1.215960	.190478	.112	1.70070	.21450	.099456	.888
.389	1.221227	.189259	.111	1.70504	.21289	.098679	.889
.390	1.226528	.188036	.110	1.70941	.21128	.097900	.890
.391	1.231864	.186806	.109	1.71382	.20966	.097119	.891
.392	1.237235	.185572	.108	1.71826	.20804	.096336	.892
.393	1.242642	.184332	.107	1.72273	.20642	.095551	.893
.394	1.248085	.183087	.106	1.72724	.20480	.094764	.894
.395	1.253565	.181836	.105	1.73177	.20317	.093975	.895
.396	1.259084	.180579	.104	1.73634	.20154	.093184	.896
.397	1.264641	.179318	.103	1.74095	.19991	.092391	.897
.398	1.270237	.178050	.102	1.74559	.19827	.091596	.898
.399	1.275874	.176777	.101	1.75027	.19664	.090799	.899
.400	1.281552	.175498	.100	1.75498	.19500	.090000	.900

.400 .100 .900

.400			.100				.900
I	*x*	*z*	*q*	*z/q*	*z/p*	*pq*	*p*
.400	1.281552	.175498	.100	1.7550	.19500	.090000	.900
.401	1.287271	.174214	.099	1.7597	.19336	.089199	.901
.402	1.293032	.172924	.098	1.7645	.19171	.088396	.902
.403	1.298837	.171628	.097	1.7694	.19006	.087591	.903
.404	1.304685	.170326	.096	1.7742	.18841	.086784	.904
.405	1.310579	.169018	.095	1.7791	.18676	.085975	.905
.406	1.316519	.167705	.094	1.7841	.18510	.085164	.906
.407	1.322505	.166385	.093	1.7891	.18345	.084351	.907
.408	1.328539	.165060	.092	1.7941	.18178	.083536	.908
.409	1.334622	.163728	.091	1.7992	.18012	.082719	.909
.410	1.340755	.162391	.090	1.8043	.17845	.081900	.910
.411	1.346939	.161047	.089	1.8095	.17678	.081079	.911
.412	1.353174	.159697	.088	1.8147	.17511	.080256	.912
.413	1.359463	.158340	.087	1.8200	.17343	.079431	.913
.414	1.365806	.156978	.086	1.8253	.17175	.078604	.914
.415	1.372204	.155609	.085	1.8307	.17006	.077775	.915
.416	1.378659	.154233	.084	1.8361	.16838	.076944	.916
.417	1.385172	.152851	.083	1.8416	.16669	.076111	.917
.418	1.391744	.151463	.082	1.8471	.16499	.075276	.918
.419	1.398377	.150068	.081	1.8527	.16329	.074439	.919
.420	1.405072	.148666	.080	1.8582	.16159	.073600	.920
.421	1.411830	.147258	.079	1.8640	.15989	.072759	.921
.422	1.418654	.145843	.078	1.8698	.15818	.071916	.922
.423	1.425544	.144420	.077	1.8756	.15647	.071071	.923
.424	1.432503	.142991	.076	1.8815	.15475	.070224	.924
.425	1.439531	.141555	.075	1.8874	.15303	.069375	.925
.426	1.446632	.140112	.074	1.8934	.15131	.068524	.926
.427	1.453806	.138662	.073	1.8995	.14958	.067671	.927
.428	1.461056	.137205	.072	1.9056	.14785	.066816	.928
.429	1.468384	.135740	.071	1.9118	.14611	.065959	.929
.430	1.475791	.134268	.070	1.9181	.14437	.065100	.930
.431	1.483280	.132788	.069	1.9245	.14263	.064239	.931
.432	1.490853	.131301	.068	1.9309	.14088	.063376	.932
.433	1.498513	.129807	.067	1.9374	.13913	.062511	.933
.434	1.506262	.128304	.066	1.9440	.13737	.061644	.934
.435	1.514102	.126794	.065	1.9507	.13561	.060775	.935
.436	1.522036	.125276	.064	1.9574	.13384	.059904	.936
.437	1.530068	.123750	.063	1.9643	.13207	.059031	.937
.438	1.538199	.122216	.062	1.9712	.13029	.058156	.938
.439	1.546433	.120674	.061	1.9783	.12851	.057279	.939
.440	1.554774	.119123	.060	1.9854	.12673	.056400	.940
.440			.060				.940

.440 .060 .940

I	x	z	q	z/q	z/p	pq	p
.440	1.554774	.119123	.060	1.9854	.12673	.056400	.940
.441	1.563224	.117564	.059	1.9926	.12494	.055519	.941
.442	1.571787	.115996	.058	1.9999	.12314	.054636	.942
.443	1.580467	.114420	.057	2.0074	.12134	.053751	.943
.444	1.589268	.112836	.056	2.0149	.11953	.052864	.944
.445	1.598193	.111242	.055	2.0226	.11772	.051975	.945
.446	1.607248	.109639	.054	2.0304	.11590	.051084	.946
.447	1.616436	.108027	.053	2.0382	.11407	.050191	.947
.448	1.625763	.106406	.052	2.0463	.11224	.049296	.948
.449	1.635234	.104776	.051	2.0544	.11041	.048399	.949
.450	1.644854	.103136	.050	2.0627	.10856	.047500	.950
.451	1.654628	.101486	.049	2.0711	.10672	.046599	.951
.452	1.664563	.099826	.048	2.0797	.10486	.045696	.952
.453	1.674665	.098157	.047	2.0884	.10300	.044791	.953
.454	1.684941	.096477	.046	2.0973	.10113	.043884	.954
.455	1.695398	.094787	.045	2.1064	.09925	.042975	.955
.456	1.706044	.093086	.044	2.1156	.09737	.042064	.956
.457	1.716886	.091375	.043	2.1250	.09548	.041151	.957
.458	1.727934	.089652	.042	2.1346	.09358	.040236	.958
.459	1.739198	.087919	.041	2.1444	.09168	.039319	.959
.460	1.750686	.086174	.040	2.1544	.08976	.038400	.960
.461	1.762410	.084417	.039	2.1645	.08784	.037479	.961
.462	1.774382	.082649	.038	2.1750	.08591	.036556	.962
.463	1.786614	.080868	.037	2.1856	.08398	.035631	.963
.464	1.799118	.079075	.036	2.1965	.08203	.034704	.964
.465	1.811911	.077270	.035	2.2077	.08007	.033775	.965
.466	1.825007	.075452	.034	2.2192	.07811	.032844	.966
.467	1.838424	.073620	.033	2.2309	.07613	.031911	.967
.468	1.852180	.071775	.032	2.2430	.07415	.030976	.968
.469	1.866296	.069915	.031	2.2553	.07215	.030039	.969
.470	1.880794	.068042	.030	2.2681	.07015	.029100	.970
.471	1.895698	.066154	.029	2.2812	.06813	.028159	.971
.472	1.911036	.064250	.028	2.2946	.06610	.027216	.972
.473	1.926837	.062332	.027	2.3086	.06406	.026271	.973
.474	1.943134	.060397	.026	2.3230	.06201	.025324	.974
.475	1.959964	.058445	.025	2.3378	.05994	.024375	.975
.476	1.977368	.056476	.024	2.3532	.05786	.023424	.976
.477	1.995393	.054490	.023	2.3691	.05577	.022471	.977
.478	2.014091	.052485	.022	2.3857	.05367	.021516	.978
.479	2.033520	.050462	.021	2.4030	.05154	.020559	.979
.480	2.053749	.048418	.020	2.4209	.04941	.019600	.980

.480 .020 .980

.480 .020 .980

I	x	z	q	z/q	z/p	pq	p
.480	2.053749	.048418	.020	2.4209	.04941	.019600	.980
.481	2.074855	.046354	.019	2.4397	.04725	.018639	.981
.482	2.096927	.044268	.018	2.4593	.04508	.017676	.982
.483	2.120072	.042160	.017	2.4800	.04289	.016711	.983
.484	2.144411	.040028	.016	2.5018	.04068	.015744	.984
.485	2.170090	.037870	.015	2.5247	.03845	.014775	.985
.486	2.197286	.035687	.014	2.5491	.03619	.013804	.986
.487	2.226211	.033475	.013	2.5750	.03392	.012831	.987
.488	2.257129	.031234	.012	2.6028	.03161	.011856	.988
.489	2.290370	.028960	.011	2.6327	.02928	.010879	.989
.490	2.326348	.026652	.010	2.665	.02692	.009900	.990
.491	2.365618	.024306	.009	2.701	.02453	.008919	.991
.492	2.408916	.021920	.008	2.740	.02210	.007936	.992
.493	2.457264	.019487	.007	2.784	.01962	.006951	.993
.494	2.512144	.017003	.006	2.834	.01711	.005964	.994
.495	2.575829	.014460	.005	2.892	.01453	.004975	.995
.496	2.652070	.011847	.004	2.962	.01189	.003984	.996
.497	2.747781	.009149	.003	3.050	.00918	.002991	.997
.498	2.878161	.006340	.002	3.170	.00635	.001996	.998
.499	3.090229	.003367	.001	3.367	.00337	.000999	.999

.499 .001 .999

INDEX

Boldface is used for references to definitions.

Alienation, coefficient of, **173**
  see also Correlation
Alignment chart of correlation functions, 291–295, *inside back end paper*
American Society of Mechanical Engineers, 42
Anderson, von, O., 271, 276
Angell, Frank, 147
Approximations, errors in, 164–167
Array, **154, 155**
Attenuation, **204–205**
Average, moving, **28**
Averages, 44–69

Bar diagram, **38**
Bell, Julia, 265
Best fit, **159**
Blakeman, John, 239, 269
Block diagram, **40**
Boas, Franz, 259
Bowley, A. L., 55
Bravais, A., 152
Bridges, Calvin B., 321
Broad categories, effect of, 167–171
Brown, Carl, 226, 227
Brown, William, 37, 47, 190, 203, 205, 326

Canning, J. W., 183
Caption, 6
Categorical measures, graphic representation of, 37–43
Cave, Beatrice M., 271, 272, 273
Cave, F. E., 271
Central tendencies, 44–69
Charlier, C. V. L., 123
Chart of ratios, **22**, 23–27
Chart, relative time, 16–20, **18**
Chart, time, **16**, 17
Charts, summary of rules for construction of, 42–43
Class index, **11–13**, 168–169
Class interval, **11, 13**
Class limits, **11, 12**
Class mean, **168–169**
Cobb, Margaret V., 314
Comparable measures, **109–122**, 153
  percentile method, 118–122
  ratio method, 110–114
  standard measure method, 114–117
Contingency
  see Correlation
Correlated measures, functions of, 196–230
Correlation, average inter-, 217–221
Correlation between
  a mean and a cell frequency, 178
  a mean and coefficient of correlation, 178
  a mean and standard deviation, 178
  any two product movements, 175
  coefficients of correlation, 179
  means, 178
  standard deviations, 178
  standard deviation and coefficient of correlation, 178
  sums or averages, 196–200
Correlation coefficient, product-moment, **161–164**
  calculation of, 179–181
  corrections to, 171
Correlation, corrected for attenuation, **204–205**
  error in, 208–212
Correlation, effect of range upon, 221–230
Correlation, interpretation of, 189–190
  graphic, 153–156
Correlation, partial and multiple, 279–310
  multiple alienation coefficient, **288, 299–300**
  multiple correlation coefficient, **287**
  multiple, three variables only, 280–295
  multiple, n variables, 283, 292, 294, 295–310
  partial alienation coefficient, 289
  partial correlation coefficient, **289, 290, 298**
  by successive approximations, 302–310

Correlation surface, normal, **156, 157–159**
Correlation surfaces, **172**
Correlation table, **154**
Correlation, various measures of, 231–278
  bi-serial eta, **249–253**
  bi-serial r, **245–249**
  contingency, 262–265
  contingency, coefficient of, **265–271**
  contingency, corrections to coefficient of, 267–271
  contingency, partial, **280**
  contingency, multiple, **280**
  equi-probable r, **265**
  four-fold point surface, 259–260
  mean square contingency, **265–271**
  non-rectilinear regression, 185–189
  Otis' rank relation, **234–237**
  parabolic regression, 245
  ϕ, *see* Correlation, four-fold point surface
  rank method, **191–194**
  ratio, correlation, **238–245**
  ratio, correlation, corrected, 241, 242, 244
  ratio, correlation, multiple, **280**
  ratio, correlation, partial, **280**
  Thorndike's median ratio coefficient, **231–234**
  tetrachoric, **253–258**
  variate difference, **271–278,** 280
  Yule's coefficient of association, **260–262**
  Yule's coefficient of colligation, **260–262**
Correlation with true measures, **200–201,** 204
Cross-over value of a chromosome section, 321–324
Curve fitting, 123–150
  normal curve, 136
  type II, 136–137
  type III, 137–138
  type V, 138
  type VII, 137
Curves, types of, **128–135**

Day, Edmund E., 2
Deviation, mean, **70–75,** 96
Deviation, quartile, 34, **75**
Deviation, 10–90 percentile range, 34, **75–77**
Deviation, standard, **77–82**
Deviation, standard, of constants of single series
  of a class frequency, 86–92
  of any moment, 84–86
  of index numbers, 334–339, 340
  of an interpercentile range, 76
  of the mean, 82–83, 177
  of the median, 90
  of measure of Kurtosis, 77
  of measure of Skewness, 77
  of a percentile, 86–92
  of 10–90 percentile range, 76
  of the standard deviation, 176
Deviation, standard, of measures of correlation
  of bi-serial eta, 250
  of bi-serial r, 249
  of coefficient of contingency, 269
  of correlation ratio, 241
  of multiple coefficient of correlation, 301–302
  of partial coefficient of correlation, 301
  of product-moment coefficient of correlation, 176
  of r corrected for attenuation, 209–210
  of r inferred from an r obtained in a different range, 316
  of rank coefficient of correlation, 194
  of regression coefficient, 176, 301, 302
  of tetrachoric coefficient of correlation, 257–258
  of ϕ, 262, 269
  of variate difference correlation coefficient, 276–277
Deviation, standard, of a difference, 182
Deviation, standard, of an array, 155, 173
  of an array mean, 177
Deviation, standard, of an estimated measure, 300
Deviation, standard, of any product moment, 175
Dickson, J. D. Hamilton, 156
Dispersion, 44, **70–93**
Duffell, J. H., 137

Edgeworth, F. Y., 123, 152, 333
Elderton, W. Palin, 47, 124, 264
Error, probable, **98**

Error, probable
see Deviation, standard
Error, standard
see Deviation, standard
Everitt, P. F., 254, 255

Fechner, G. T., 326, 327
Filon, L. N. G., 176, 179
Fisher, Irving, 332, 333, 335, 339, 341, 342
Forsyth evaluation of the Gamma function, 136
Frequency polygon, 9, **10**, 11–15

Galton, Francis, 114, 152, 153, 155
Gauss, C. F., 153
Graphic methods, 9–43
Greek alphabet, 356
Grouping, **50**, 167
Grouping, effect upon correlation, 167–171
Grouping, rule for, **52**
Grove, C. C., Preface, p. vi
Growth curve, 34
Growth increments, 35–37

Haskell, Allen, C., 42
Heron, David, 259, 261, 262, 266
Herring, John P., 72
Histogram, 9, **10**, 11–12
Holzinger, Karl J., 222
Homoclisy, **172**
Homoscedasticity, **172**
Hooker, R. H., 271

Indexes, 66–67
Index numbers, **331–347**
change of base of, 346–347
flexibility of geometric, 339–341
meaning of, 333–334
tests of, 341–346
Isserlis, L., 179, 188

Kapteyn, J. C., 123
Kelley, Lura, 97
Kelley, Truman L., 75, 90, 97, 115, 173, 213, 221, 223, 291, 297, 298, 321, 333, 338, 342
Kelley-Wood table, 97, 370–385
Knibbs, G. H., 333
Kurtosis, **45**, 77

Labelling classes, 53
Lee, Alice, 255, 311
Lengthening tests, effect of, 205–208

Map diagram, **39–41**
Mean
arithmetic, 44, **45–53**
geometric, 65–**66**
guessed, **48**
harmonic, 63–**64**
Median, 34, **54–57**
Mitchell, Wesley C., 333, 339
Mode, 34, **60–62**
Moments, 48, **79**
Müller, G. E., 327
Mutilated distributions, constants of, 311–314
Mutilated distributions, correlation in, 314–316

Normal curve, fitting a, 136
Normal distribution, **94–108**, 129–130, 145, 149
Normal distribution, unit, **99–100**, 350

Ogive, **31–34**
Origin, arbitrary, **48**
Otis, Arthur S., 118, 234, 237
Overlapping, error in measures of, 213, 316–319

Pearson, Karl, Preface, pp. v and vi, 94, 99, 123, 124, 125, 135, 137, 138, 140, 141, 143, 152, 153, 169, 172, 174, 175, 176, 179, 193, 194, 225, 229, 231, 239, 241, 248, 249, 250, 253, 254, 257, 259, 261, 262, 264, 265, 266, 268, 269, 271, 272, 273, 311, 333, 336
Percentiles, 34, **57–59**
Perry, C. A., 39
Persons, Warren M., 271
Pintner, Rudolph, 115
Population, 44
Probability of exceeding a given divergence, **102–103**
Probable error, **98**
Probable error
see Deviation, standard
Probable error of estimate
see Deviation, standard, of an array
Product theorem in correlation, 84

Product theorem in probabilities, **262**
Psychophysical methods, 326–330.

Ratios, 66–67, 110–114
Regression, **152, 154**
  see also Correlation
Regression coefficients, **160–161,** 181–185
  conjugate, **298**
  3 variables, 283, **285–294**
  n variables, **283,** 292, 294, 295, 296–302
Regression equation, **161**
  3-variables, **281, 283**
  n-variables, 283, **295–310**
Relationship, measures of, 151–195
  see Correlation
Reliability coefficient, **200–203**
Residual, definition of, **281,** 284
Reversion, **152, 154**
Rhind, A., 143
Rich, Willis H., 273
Richmond, H. A., 185
Rietz, H. L., Preface, p. vi, 189
Ritchie-Scott, A., 271
Rugg, Harold O., 39

Scatter diagram
  see Correlation table
Series, statistical, 2–5
  complex, 39–41
  qualitative, 2, **5**
  quantitative, 2, **5**
  spacial and geographical, 2, **4**
  temporal, 2–3
Sheppard, W. F., 94, 125, 168, 169, 174, 176, 257
Similar forms, **201–203**
Skewness, **44,** 77
Smoothing data, **27–31**
Soper, H. E., 177, 248, 249
Spearman, Charles, 193, 196, 203–204, 205, 210
Stability of distributions, 138–150
Standard error, **83**
Standard error of estimate
  see Deviation, standard, of an array
Standard measures, **115, 280**
Stub, **6**
Student, 243, 271, 272
Symbols, list of important, 349–356

Tables, statistical, **5–8**
  derived, **7**
  general purpose, **7**
  primary, **7**
  special purpose, **7**
Thiele, T. N., 123
Thomson, Godfrey H., 37, 47, 190, 203, 326, 327
Thorndike, E. L., Preface, p. vi, 231, 234
Thurstone, L. L., 199
Trade test evaluation, 320–321
True scores, **200**
  estimates of, 212–216
  error in estimates of, 212–216

Unit normal distribution, **99–100**
Unstable distributions, 146–150
Urban, F. M., 327, 328

Variability
  see Dispersion

Wald, Elva, 72
Walsh, C. M., 333
Weighted average, best, 324–325
Weighting, **67–68**
Weighting, effect of, 199–200
Weightings, merit of fixt, 319–320
Weldon, W. F. R., 189
Whipple, G. M., 111
Wood, Ben D., 97
Woodyard, Ella, 309

Yerkes, Robert M., 226, 228, 246, 250, 255, 269, 309
Yule, G. U., 160, 210, 259, 260, 261, 262

LIBRARY
UNIVERSITY OF MIAMI

LIBRARY
UNIVERSITY OF MIAMI

k
r_{13}
Product
r_{23}
R
1000
900
800
700
600
500
400
300
200
150
100
950
940
930
920
910
850
700
600
500
400
300
200

Richter HA29.K4
Statistical method
3 5051 00566 6070

WITHDRAWN
From UF Library

Date Due

12Apr '51	~~9 Mar 69 Y~~		
3 Apr '52	~~2 May '69 W~~		
Jan 30 53			
Oct 29 '53	~~JUL 7 1976~~	~~CO-RC~~	
Apr 22 '54	116-2		
29 Aug 57 M			
~~NOV 22 1958~~			
~~15 Oct '62 M~~			
~~26 Apr '63 F X~~			
~~28 Apr '63 Y X~~			
~~23 Oct '63 W~~			
~~18 Dec '63 W~~			
~~13 May '64 W~~			
20 Jul 66 W			
17 Jan 69 W			